UNECE

Geospatial information services based on official statistics

Working Paper Series on Statistics

Issue 1

UNITED NATIONS

UNECE Working Paper Series on Statistics, Issue 1, October 2016

In-depth review of developing geospatial information services based on official statistics

United Kingdom Office for National Statistics[1]

Abstract

This review is concerned with how geospatial information can be used in the production of official statistics. It will consider how geospatial information fits within the statistical process and what barriers exist to exploiting the value of geospatial information for official statistics.

[1] Note written by Ian Coady, Office for National Statistics. Additional contributions by Jerker Mostrom (Statistics Sweden), Ekkehard Petri (Eurostat), Denise McKenzie (Open Geospatial Consortium), Neil Sutherland (Ordnance Survey), Marja Tammilehto-Luode (Statistics Finland), Martin Brady (Australian Bureau of Statistics), Gemma van Halderen (Australian Bureau of Statistics), Keith Spicer (Office for National Statistics), Bill Oates (Office for National Statistics) and Nigel Swier (Office for National Statistics)

Table of Contents

I. Executive summary

1. The need for accurate and reliable statistics has never been more apparent. The United Nations Sustainable Development Goals (SDGs) have added a new set of targets that will need to be monitored through official statistics. As the world faces complex challenges to manage limited resources there will also be an increasing need for statistics to inform how, when and where those resources should be allocated.

2. National Statistical Institutes (NSIs) have to be increasingly flexible to be able to deliver the range of statistics required. This flexibility is coming through the modernisation of official statistics which UNECE is facilitating through the Common Statistical Production Architecture and Generic Statistical Business Process Model (GSBPM). These are flexible frameworks that should allow geospatial information to be integrated into the statistical process without too much impact on the existing organisational structure.

3. The reason that more geospatial information is needed within the statistical process is that statistics increasingly need to be understood within the context of location. Statistics have always been published on the basis of geography but it is now important to understand what area that geography represents and what additional context the geography can add to the statistical data. To do this NSIs need access to a wide range of geospatial services from the ability to collect and georeference data at the address point level, to the geographic zoning tools to define enumeration areas and geographies for dissemination that can improve the quality of statistics and reduce the risk of disclosure by differencing.

4. To deliver this closer integration of statistics and geospatial information, the United Nations established its Committee of Experts on Global Geospatial Information Management (UN-GGIM) to undertake a global consultation programme on geocoding, linking and integration, and geographic classification practices related to statistics.

5. The best way to do this is through the development of a Global Statistical Spatial Framework that will set out generic guidance and best practice for the integration of statistics and geospatial information. The UN-GGIM Expert Group on the Integration of Statistics and Geospatial Information has been working to adapt the Australian Bureau of Statistics' Statistical Spatial Framework to incorporate additional elements that will provide this generic framework.

6. The modernisation of official statistics is facilitating the integration of new data sources such as administrative and big data within the statistical system but this presents a new set of challenges for the use of geospatial data.

7. Data are increasingly needed at small geographic levels to support initiatives such as the post-2015 development agenda and the Sustainable Development Goals, UN-GGIM and the GEOSTAT 2 project are providing the frameworks for this, but administrative and big data are can often only be aggregated to higher geographic levels due to disclosure and quality concerns.

8. There is therefore a mismatch between the increased use of point-based data for the aggregation of statistics, and the availability of some datasets at larger geographic areas than those needed to monitor sustainable development. Increasingly NSIs will need to look at disaggregation as a methodology for providing statistics for monitoring. The implications of these disaggregation methodologies on official statistics will need to be fully explored through UN-GGIM and other initiatives.

9. Increasing the geospatial content within the statistical process also increases the risk of disclosure. Slivers generated by overlapping geographies, the ability to link statistical attributes to a single grid reference and the need to produce statistics for smaller geographical areas increase the risk of being able to identify sensitive data. NSIs have worked to develop disclosure control methodologies that will compensate for these risks

but the introduction of new data sources and changes to the statistical process are likely to again increase the risk of disclosure.

10. An increased use of technical and non-technical standards is helping to deliver a closer integration of statistics and geospatial information. However, statistical standards and geospatial standards still sit within their own domains and there is little current work on the harmonisation of statistical and geospatial standards. This gap has been recognised by the Open Geospatial Consortium and the International Standards Organisation. The statistical community will need to consider how best to engage in this work.

11. All of this work is to move towards the integration of more geospatial data into the statistical process. The objective must therefore be for statisticians to have a greater understanding of the value of geospatial information and to see that understanding manifested as an increase in spatial analysis alongside statistical analysis within reporting. All of this should support the modernisation of official statistics as well as add value to the statistical data being produced by NSIs.

II. Introduction

12. The Bureau of the Conference of European Statisticians (CES) reviews regularly selected statistical areas in depth to improve coordination of statistical activities in the United Nations Economic Commission for Europe (UNECE) region, identify gaps or duplication of work, and address emerging issues. The review focuses on strategic issues and highlights concerns of statistical offices of both a conceptual and a coordinating nature. The review is based on a paper prepared by an invited country/organisation. The paper provides the basis for a review of activities related to the development of services through the integration of statistics and geospatial information by summarising international statistical activity in this area, identifying issues and problems and making recommendations on how these issues should be followed up.

13. At the February meeting of CES Bureau, the topic of developing geospatial information services based on official statistics was selected for an in-depth review at the February 2016 meeting of the Bureau.

III. Scope/definition of the statistical area covered

14. This review is primarily concerned with how geospatial information can be used in the production of official statistics. It will consider how geospatial information fits within the statistical process and what barriers exist to exploiting the value of geospatial information for official statistics.

15. In the context of this review, geospatial information services are the geographic components of the statistical process needed for the delivery of official statistics.

16. For this review, official statistics are considered to be statistics published by government organizations to provide quantitative or qualitative information on areas such as society, economy, health, education or the environment.

17. The review will focus on existing work within the international statistical community to harmonize the integration of statistics and geospatial information and will consider how this work fits into existing statistical frameworks implemented by the UNECE. The review has not considered any topic to be out of scope.

IV. Overview of international statistical activities in the area

A. Link with GSBPM and CSPA

18. A number of activities have already taken place to harmonise the statistical process at the international level. These activities have been driven by the identification of two major threats to the supply of core statistics the UNECE countries. These are: (i) rigid processes and methods, and (ii) inflexible ageing technological environment.

19. To mitigate the risks posed by these threats to production of statistics UNECE has coordinated the work to develop two frameworks for the modernisation of official statistics that will provide flexibility to processes and technology.

20. The Generic Statistical Business Process Model (GSBPM) defines the business processes needed to produce official statistics as well as providing a framework for quality assessment of data and processes.

21. GSBPM should be interpreted flexibly rather than as a rigid framework in which all steps must be followed in a strict order. It instead focuses on the relationships and dependencies between processes. By taking a matrix approach to the statistical process, statistical producers are able to assess the most appropriate path for statistical production rather than adhering to the rigid processes that has been identified as a threat.

22. The structure of GSBPM comprises three levels, the statistical business process, the eight phases of the statistical business process and the sub-processes within each phase. For the purposes of this review, geographic activities will be matched against the eight phases of the statistical business process to describe how they might fit into the modernisation of official statistics.

23. The eight phases are:

i. Specify needs

ii. Design

iii. Build

iv. Collect

v. Process

vi. Analyse

vii. Disseminate

viii. Evaluate

24. Eurostat's GEOSTAT 2 project has already started to evaluate GSBPM from the viewpoint of using geospatial data in the production of statistics. Geospatial activities within the statistical process were identified first, followed by a search of the corresponding GSBPM phases. The project is still in progress but common characteristics of geospatial data use in statistical production processes have already being identified.

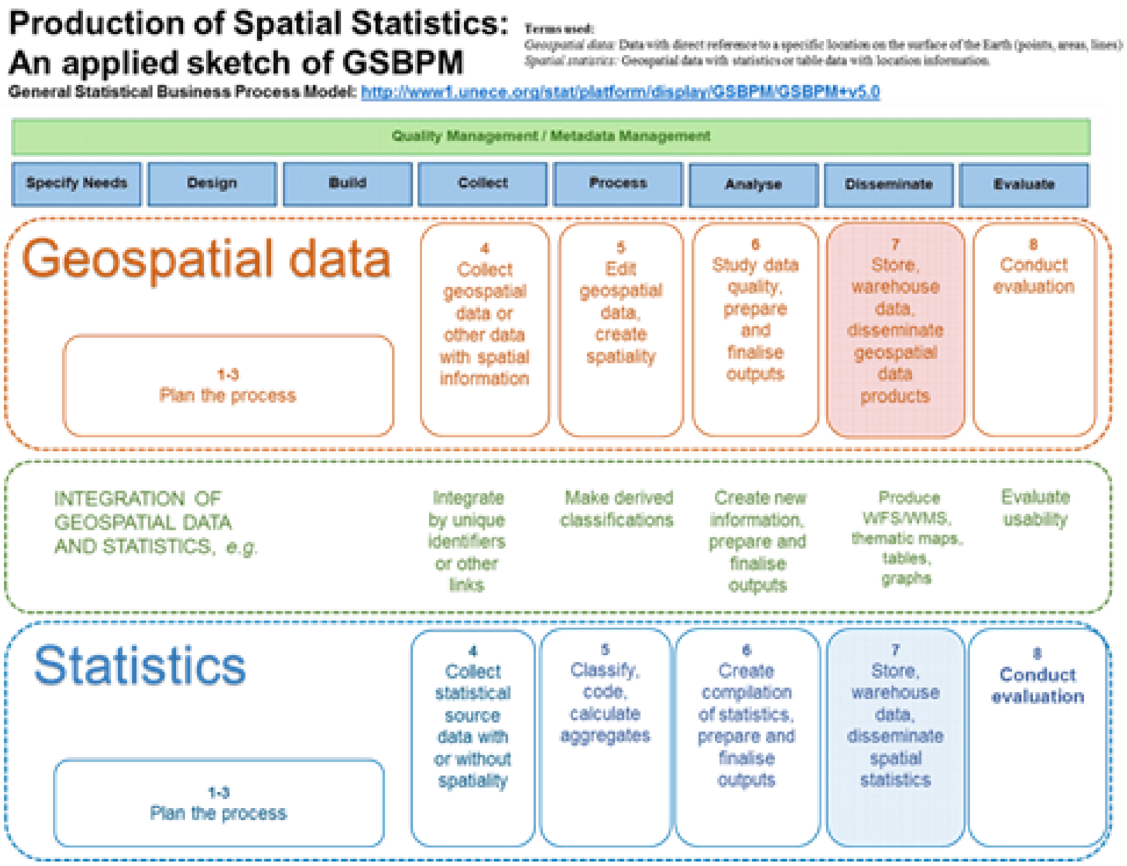
25. Geospatial data may be involved in the statistical production process from the beginning until the end. It may also only be a part of one particular phase. The role of geospatial data varies. It may be source data and a starting point for production. It may also be the outcome and end product of the process. The role of geospatial data may also be in the form of auxiliary data used supplement statistical data.

26. The production of geospatial statistics seems to fit to GSBPM. However in order to recognise GSBPM phases, plenty of interpretation is needed first. The GSBPM guidelines

do not include any reference or description about using geospatial data therefore activity is currently focussed on defining the geospatial processes within the GSBPM rather than any implementation. The GEOSTAT 2 project aims to draft an additional descriptive text for GSBPM about the spatial dimension of statistical production.

27. The GEOSTAT 2 project believes that GSBPM[2] can be used to promote common concepts and unified methods for the production of geospatial statistics. It may increase mutual understanding between parties involved by giving a common way to communicate. Best practices, instructions and tools can be better shared. This kind of communication may help NSIs recognise similar phases from different processes and make them more coherent.

28. In addition to GSBPM, the High Level Group for the Modernization of Statistical Production and Services (HLG) has put priority on the development of a Common Statistical Production Architecture (CSPA) as a generic architecture for statistical production. By adopting this common reference architecture, it will be easier for organizations to standardize and combine components of statistical production.



29. The CSPA differs from the GSBPM in that it is not a defined statistical process but rather an approach to the statistical architecture needed to implement that process. The goal of CSPA is to provide statistical organizations with a standard framework to facilitate the modernization of official statistics and provide guidance on how that modernization should be implemented within statistical organizations. The CSPA also considers how it can align with the GSBPM to best deliver a coordinated approach to statistical modernization.

---

[2] For further information, check http://www1.unece.org/stat/platform/display/metis/The+Generic+Statistical+Business+Process+Model

30. This review will set out the overview of statistical activities and the country practices in the context of the CSPA[3] as a framework for implementing change to help understand what impact implementing those changes could have on the architecture of statistical organisations. It will also consider where those changes might fit into the GSBPM at a high level.

B. Drivers for international integration of statistics and geospatial information

31. There are a number of drivers for closer integration of statistics and geospatial information. The requirement to develop an integrated statistical and geospatial solution to the 2020 round of population censuses has been repeatedly expressed in the UN context, e.g. by the UN Committee of Experts on Global Geospatial Information Management (UN-GGIM) in its report from 2013 and the UN Expert Group on the Integration of Statistics and Geospatial Information (EG-ISGI).

32. Within the National and International Statistical Systems there is a move towards an increased use of administrative data and registers for census purposes. In parallel, many countries have launched national geospatial strategies to georeference administrative records to support data linkage. Official Statistics can take advantage of this increased data linkage through geocoded administrative data for the 2020 round of population censuses. This has been recognised by the ESSC in its 19th meeting in November 2013: "The newly created Task Force was supported and requested to develop a strategy for a harmonised approach to geo-referenced statistics within the ESS and to base the 2021 round of censuses on registers."

33. The ESS Task Force addressed this issue by recommending a comprehensive geocoding strategy for the next census[4]. All census information should be georeferenced to a single point that represents an address, building or dwelling from a national register.

34. It should be noted that not only register based censuses can benefit from such a point based spatial reference framework. Traditional censuses, if based on the same type of spatial reference system would allow for the same type of geospatial statistical outputs, in particular dissemination on small area geographies.

35. For traditional censuses, also the preparation of the enumeration and the management of the actual enumeration would benefit from geospatial information support, notably the preparation and execution of the enumeration campaign in the field.

36. This proposal is taken up in the UNECE census recommendations to associate the place of usual residence, and the place of employment with a pair of coordinates to a point. This creates the necessary infrastructure to produce population grids.

37. The European Statistical System is currently working on a legal base for the production of in total 13 different 1km² population grids, mainly recording key demographic and residential characteristics of an individual person:

- Total population
- Sex (males, females)
- Age (under 15, 15-64, 65 and over)

---

[3] For further information, check http://www1.unece.org/stat/platform/display/CSPA/Common+Statistical+Production+Architecture+Home

[4] https://circabc.europa.eu/d/a/workspace/SpacesStore/fd349927-3c7d-435a-8b80-4ace48daf646/D_GIS_105%20GISCO-TF-Report-V_3.doc

- Employed persons
- Place of Birth (in the reporting country, in another EU country, outside EU)
- Usual residence 12 months before (unchanged, within reporting country, outside of the reporting country).

38. The UNECE recommendations go one step further and also recommend associating the place of work or education to a point. This would allow constructing day-and-night-time information at high resolution and also supporting commuting statistics that is difficult to create from register based-censuses.

39. The need for point based data collection and geographic enumeration of census data will impact on the 'specify needs' and 'design' phases of the GSBPM.

40. The proposal for a corresponding point based spatial reference framework based on geocoded address, building and dwelling registers is the main subject of the GEOSTAT 2 project.

C. Statistical Spatial Framework

41. At the forty-fourth session of the United Nations Statistical Commission, held in February 2013, the Statistical Commission discussed the Programme Review: "Developing a statistical-spatial framework in national statistical systems".

42. The UN-GGIM Committee of Experts established an Expert Group to carry out the work on developing a statistical-geospatial framework as a global standard for the integration of statistical and geospatial information.

43. Also within the specification of needs and statistical design is the need for a global statistical spatial framework. This framework will provide policy, guidance and best practice on how geospatial information integrates into every phase of the GSBPM but will particularly impact the design of statistical production as producers will need to set out how geospatial information in going to be integrated from the start.

44. One of the key deliverables of the Expert Group was to "evaluate the statistical-geospatial framework developed by the Australian Bureau of Statistics, and determine if and how this could be internationalised".

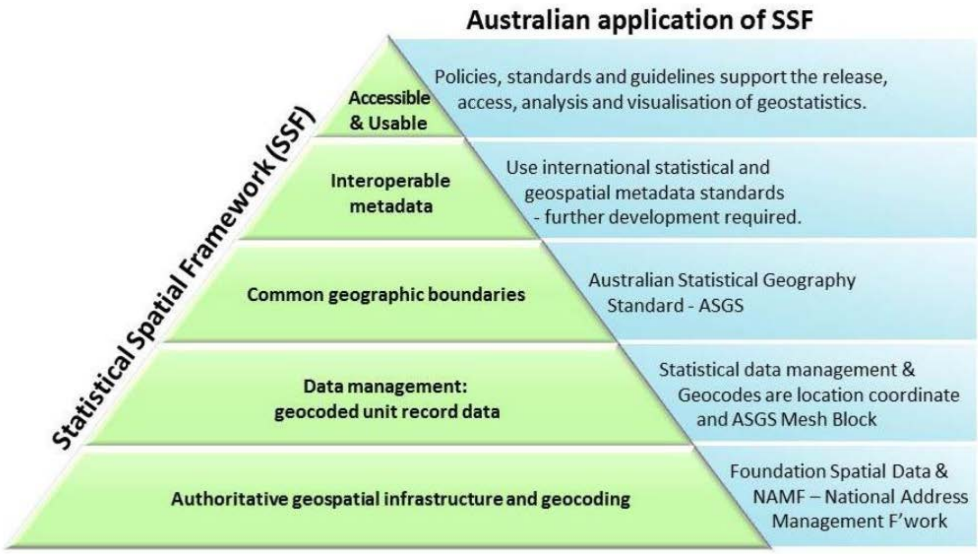
45. The Statistical Spatial Framework developed by the Australian Bureau of Statistics (ABS) is a response to the challenge of better integration of geospatial and statistical information which has large overlaps with the objectives of UN-GGIM. This framework provides Australia with a common approach to connecting socio-economic statistical data to a location, and improves the accessibility and usability of this location-enabled information.

46. The generic Statistical Spatial Framework, developed by the ABS, consists of five elements that are considered essential to integrating geospatial and socio-economic information (see green layers in diagram below). The Statistical Spatial Framework for Australia details the Australian implementation of this generic Framework (see blue layers in diagram below).

47. The Expert Group has sought to expand the work of the ABS Statistical Spatial Framework by incorporating elements from the Mexican Statistical Institute's approach to geospatial statistics and the work by Statistics Finland and the GEOSTAT team to expand the Generic Statistical Business Process Model to include geospatial information as described in section 7 of this review.

48. When combined these three approaches should form a Global Statistical Spatial Framework that is generic and flexible enough to be applied by all NSIs. The UK has

offered to use its own statistical framework, the GSS Geography policy to help quality assure the Global Statistical Spatial Framework.



D. Core data

49. The requirement to integrate statistics with geospatial information has more firmly established the requirement of National Statistical Institutes (NSIs) to have access to key geographic datasets that will be needed for statistical production for the collection, processing, analysis and dissemination phases of the GSBPM.

50. The conventional separation between NSIs and national mapping & cadastral agencies (NMCAs) has, in the past, hampered the development of interoperability between the two disciplines. There are, however, a number of initiatives aimed at ensuring that the geography and statistics communities will work more closely together in future.

51. At its first meeting in October-November 2013 the UN Expert Group on the Integration of Statistics and Geospatial Information agreed to undertake a global consultation programme on geocoding practices, linking and integration practices, and geographic classification practices. The aim was to better understand and articulate practices, particularly a comparative analysis of grid-based versus population-based approaches to geocoding, integration and geographic classifications. The merits of these two approaches were also to be explored in a separate technical research programme. The Expert Group accepted the need to focus initially on the 2020 round of population censuses, while acknowledging that other statistical themes would eventually fall within scope.

52. UN-GGIM has established a series of five regional committees, each focusing on issues of particular relevance to its own region. One of these, UN-GGIM: Europe, has decided to address the integration of geospatial and statistical information and has established two work groups, membership of which is open to both NMCAs and NSIs:

- UN-GGIM: Europe Work Group A – Core Data
- UN-GGIM: Europe Work Group B – Data Integration

53. The initial focus of Work Group A has been on defining a set of 'core reference data' needed for effective geospatial data management to support the UN Sustainable Development Goals. On the basis of this work the fifth session of the UN-GGIM Committee of Experts, held in August 2015, decided that UN-GGIM: Europe should liaise with the other Regional Committees of UN-GGIM to bring together information on the

consideration of fundamental geospatial data themes and to lead discussions between interested Member States and observers in order to develop agreement of a minimum set of global fundamental geospatial data themes. At the time of writing it remains to be seen how the NSIs may be engaged in this work and whether key geospatial data that are particularly relevant for statistics will be represented among the data themes to be agreed.

54. Work Group B's initial phase of work has concentrated on the definition of priority user needs for combinations of data. The task is divided into two parts: first to define what combinations of data provide the greatest value to users, with a particular focus on the UN Sustainable Development Goals, Digital Agenda for Europe and Europe 2020 targets; and second to define which types of geospatial data, especially core data, and which types of statistical and other thematic data, should be given priority. The work has been progressed through a series of use cases. The most important recommendation to date is for all countries in Europe to create a spatial reference framework for statistics, with geocoded administrative address, building and dwelling registers at its heart.

55. ECOSOC will review the progress made by UN-GGIM in April 2016, and is expected to decide on the future form and funding of the initiative. This decision will have a substantial bearing on the prospects for global agreement on the integration of geospatial and statistical data, and on the likely timescales for progress towards this goal.

E. Geocoding

56. At the collection phase of the GSBPM it is increasingly important for NSIs to have access to authoritative address, building or dwelling registers that can be used as the basis for providing more qualified descriptions and analyses of society, economy and environment.

57. A task force led by Eurostat, the statistical office of the European Union, on the integration of statistical and geospatial information noted, inter alia, that:

- Integration is more than the combination of final information products (geospatial and statistical). The benefits of integration should be investigated and exploited during all stages of the statistical production process

- Geocoding of statistical and administrative data at unit record level is the most important

58. There is an increasing awareness that traditional surveys and censuses with fixed output areas, such as enumeration districts or other small areas, do not meet the user demands on territorial flexibility. Building the production of geospatial statistics on a spine of geocoded address information and/or building information, is believed to increase the flexibility of the output data as well as increase the spatial resolution of the data. A production system with single coordinate points as the spatial key-element will be able to deliver aggregations at any spatial unit, allowing NSIs to build their own portfolios of small areas statistics, as well as supporting on-demand aggregations on user-defined geographies.

59. Another advantage of a point-based production system is use of non-aggregated data to enhance spatial analyses within the NSIs. This is particularly relevant for accessibility studies were non-aggregated data obviously can provide more accurate output than aggregated data (e.g. gridded statistics). In the wake of climate change, scenarios for flooding and other natural hazards such as storms and fires, has become important for crisis management. In such studies, point-based location of people and dwellings increases the relevance of geocoded statistics.

60. In February 2015, Eurostat and the European Forum for Geography and Statistics (EFGS) launched GEOSTAT 2 – a two year ESSnet grant project to support the increased use of point data in the statistical process. The aim of the project is to deliver a generic

model for a geospatial reference framework for geo-enabling statistics, applicable for the ESS countries. Already from the start of the project, the idea of building such a model on a point-based approach has been in focus along with the idea that the proposed reference framework should look beyond population statistics and be suitable for statistics in the widest possible sense.

61. In order to find out about the various conditions in the ESS, regarding use of point based data for geocoding statistics, the GEOSTAT 2 project carried out a survey to NSIs in EU member states, candidate countries and potential candidates. The response indicates a strong development in many countries in recent years, towards point-based production of statistics.

62. The current situation regarding existence of and access to, data to spatialize statistics is far better than expected before launching the survey. A promising indicator is the high (and growing) share of countries in possession of address registers and building registers, which are considered to be core features of national location frameworks. More than 90 percent of the responding countries report that they have geo-enabled address information covering parts or whole of their territories. Far more than half of the responding countries report coverage of 90-100 percent of the territory. The situation for building information is quite similar. However, it should be noted that the existence of address information and building information does not necessarily imply that the information is used for geocoding purposes. When asked if geocoding of statistical information or administrative data has actually been undertaken, roughly 80 percent responded positive.

63. When asked about the prospect of building the next population census 2021 on a point-based foundation, 75 percent of the ESS countries either responded that they already have, or foresee to have, the necessary data and infrastructure in place to geocode the census data to the level of single coordinates. Roughly 10 percent reported that they probably or possibly will. The remaining countries either responded negative or did not respond at all.

64. Despite a promising development in general, there are also serious obstacles to overcome. According to the GEOSTAT 2 survey, the main reasons preventing countries from conducting geocoding are lack of resources and lack of knowledge. The second most prominent reason reported, is lack of geospatial data or that geospatial data have poor quality.

65. Though the goal is clear, there is no one-size-fits-all strategy to get there. Shifting from an area-based production system to a point-based one is not an easy operation. The technical aspect is one side to it, but also legal aspects, resources and knowledge, traditions and last but not least cooperation with producers of geospatial information is of significant importance.

66. Typically collection and maintenance of address information and building information is under the responsibility of Mapping and Cadastral Authorities. For various reasons, in some countries, access to this kind of information is restricted, too expensive or of poor quality. The set-up of authoritative location frameworks, such as address information, in a country is typically imposed by broader strategies to improve provision of services and security to citizens (postal services, police, health care, fire service etc). Even though the NSIs themselves may have limited impact on decisions to set up this kind of location frameworks, active participation of the statistical community can put forth additional arguments for implementation of authoritative location data.

67. A point-based production system cannot run on the presence of coordinates alone. There needs to be unique identifiers and consistent keys linking statistical information or administrative data to location. In terms of traditional collection and structuring of statistical information, e.g. census data, it is within the scope of the statistical institutes themselves to support implementation of keys compliant with those used by authoritative

location data (e.g. standardised address codes). Increased use of administrative sources poses a greater challenge in terms of implementation of unique identifiers as those data sources to a large extent are external to NSI's.

68. The GEOSTAT 2 survey indicates that in less than one third of the countries, there is clearly a formal policy or custom among public institutions to use standardised identifiers in registers or administrative records (address information, personal IDs, real estate codes, building IDs etc.). Not surprisingly, this category is dominated by countries with a long-standing tradition with centralised registers, such as the Nordic countries, Austria and the Netherlands. In some forty percent of the countries, such policies exist, however only to a limited extent.

69. In order to create better conditions for geocoding of administrative data, public stakeholders at all level of government and administration, need to agree on the beneficial use of authoritative location data to geo-enable all public data. Obviously, such decisions are not in the hands of the NSIs and will require political and legal support, unless it is not already in place.

F. Big Data

70. Big data is becoming widely used within the private sector to derive commercial insights. The official statistics community has started to engage in the world of big data, but remain some distance behind. A concerted push is needed to understand what big data means for official statistics and policy-making.

71. UNECE defines big data as "data that is difficult to collect, store or process within the conventional systems of statistical organizations. Either, their volume, velocity, structure or variety requires the adoption of new statistical software processing techniques and/or IT infrastructure to enable cost-effective insights to be made."[5]

72. There are numerous challenges to be overcome for incorporating big data into official statistics, including access, legal and ethical issues, technical challenges, and developing capability. Another major challenge is in understanding the trade-off between the near real-time and highly granular insights offered by big data against slower, less granular, but more robust survey-based methods with measurable error characteristics. The future of official statistics could be one which combines existing expertise on surveys with big and administrative data in a way that exploits the strengths of all available sources.

73. If this is indeed the future of official statistics, there will be a place for existing statistical frameworks. Indeed, frameworks may become even more important as a way bringing together and making sense of disparate sources. The role of geography may become particularly important and there are some early indications of this in work already being undertaken within the Office for National Statistics (ONS).

74. Research has been completed into clustering geo-located data from Twitter to infer residence and mobility patterns[6]. This involved developing a method of processing these data so it can be compared with sub-national population estimates. Another step in the methodology involved linking clusters to address data using nearest neighbour methods and

[5] UNECE "How big is big data? Exploring the Role of Big Data in Official Statistics", http://www1.unece.org/stat/platform/download/attachments/99484307/Virtual%20Sprint%20Big%20Data%20paper.docx?version=1&modificationDate=1395217470975&api=v2, March 2014, Accessed 04 December 2015

[6] Using geolocated Twitter traces to infer residency and mobility, GSS Methodology Series No.41 http://www.ons.gov.uk/ons/guide-method/method-quality/specific/gss-methodology-series/gss-methodology-series--41--using-geolocated-twitter-traces-to-infer-residence-and-mobility.pdf, Accessed 04 December 2015

using the characteristics of the address to identify which clusters are in residential locations. This illustrates the potential of address data and other geography frameworks to add value to big data.

75. Relating big data sources to geography using this type of approach could also open the door to the use of big data for improved model-based small areas estimates. We know that many big data sources are imperfect and often not representative of the population. However, model-based approaches could enable big data to be combined with other sources to identify statistical relationships based on spatial units. By borrowing strength from big data sources it should be possible to improve existing model based estimates and provide more granular statistics.

G. Disclosure Control

76. Releasing information from socio-economic datasets helps to realise the value of that dataset by:

- providing information for business processes and management,
- informing and engaging the public,
- promoting innovation and economic growth, and
- realizing a commercial return.

77. Release of information from a dataset must also comply with the provisions of national data protection legislation within each country producing statistics. The inconsistency between the requirements of different countries for data protection makes it difficult to have a harmonised approach to disclosure control.

78. Release of information may also need to take into account ethical considerations, both professional and general. For some organisations and government agencies, release of information may also need to comply with organisational polices and/or government policies.

79. To protect the privacy of an individual's or organisation's information in socio-economic datasets, the data being released may need to be de-identified or confidentialised. This will prevent an individual or organisation from being identified within the data and the private information they have provided from being disclosed.

80. De-identifying or anonymising data removes or modifies some or all of the identifying features, such as name, address, date of birth, and gender.

81. Confidentialising data seeks to minimise the risks of identification of an individual or organisation through the presence of very rare characteristics or the combination of unique or remarkable characteristics by applying some or all of the following methods:

- removal of directly identifying features or characteristics;
- careful design of the format of the released data; and/or
- modification to the content of the data released.

82. When data is released with a location or region component privacy risks can be heightened. The number of characteristics that are needed to uniquely identify an individual or organisation generally decreases as the size of the region diminishes. When data is released for point-based locations (i.e. with coordinates) identification is a relatively simple task.

83. The geography level at which tables and microdata can be released serves as an initial protection against disclosure. The threshold at which bivariate tables is typically set

at a geography of around 100-300 persons or 40-120 households, though there can sometimes be population or household counts available at lower geography. For instance, the New Zealand threshold is set at 'meshblocks', which average 100 people, while UK output areas are at least 100 persons and 40 households (average size is 300 persons and 125 households).

84. In order to address the risks of identification and consequent disclosure of attributes of individual households and their component individual persons, UK have employed record swapping in their national population census. This entails swapping households between different geographic areas, targeting those households that are more likely to be identified from unusual characteristics. Records selected for swapping will be matched on various household characteristics, such as household size. Generally speaking, swapping is kept to a low level and not all unique records are swapped. The principle is to rely on sufficient uncertainty in any disclosure claim. United States have also used record swapping, considering uniqueness of combinations of values in key variables, targeting and matching.

85. Many countries have concentrated efforts on protecting tables rather than the underlying microdata. The leaders in the previous round of censuses have been Australia, with cell counts in tables perturbed using a cell key method and a look-up table to aid consistency between tables. At the Joint UNECE/Eurostat work session on statistical data confidentiality in Helsinki in October 2014, the presentations and discussions showed that a number of countries are using or designing methods based on the Australian (ABS) method, not just for censuses but for other data collections.

86. The cell perturbation method is central to flexible table generation, where a user may request a table by selecting variables from a prescribed list. This request is then dealt with, either by an assessor or by an automated process, and the table request accepted or rejected. Australia has been the lead in this area and UK are currently working on the viability of such a method for the 2021 UK Census.

87. Further protection is afforded in tables by table design and/or by rounding cell counts where the average cell size is lower than or equal to a set threshold (two per cell in New Zealand). This aids in protection of very unusual people who may be apparent in sparse tables with a large proportion of zeros or small counts.

88. Microdata are inherently more risky due to the number of variables in a record. The underlying disclosure risk is that an intruder will make an identification of an individual based on a subset of those variables, and then discover / disclose all other information about the same individual. The protection is based around licensing and access conditions, the detail in the constituent variables, of which the level of geography is the most effective in protecting individual records. Netherlands use a rule that a region that can be distinguished in microdata on individuals must have at least 10,000 inhabitants. In UK the level of geography is broadly set at country or region for public use files (there are ten regions, and the smallest has a population of 2-3 million individuals), while those available in safe settings or data laboratories may be at local authority level (that may be as small as around 35,000 population) with some geodemographic classifications in addition.

H. Spatial Analysis

89. Analysis in the geospatial context is increasingly required to understand not only the mathematical distribution of the data but also the spatial distribution.

90. The drivers for the production of new statistics such as the United Nations Sustainable Development Goals (SDGs) provide a new set of analysis requirements to be able to drive the policy decisions needed for the post-2015 development agenda. Geospatial statistics, in a very broad sense, can make several important contributions to the proposed

indicator framework for monitoring progress towards achievement of the SDGs. Their contribution can be considered in at least three ways. Firstly, contributing directly to the proposed indicator framework; secondly, through globally accepted standards and frameworks to support coherence and international comparability; and thirdly, through innovation and modernisation.

91. The outcome document of the United Nations Summit for the adoption of the post-2015 development agenda[7] called for nations to "exploit the contribution to be made by a wide range of data, including Earth observation and geospatial information, while ensuring national ownership in supporting and tracking progress.". Earth observations, geospatial data and geospatial information have a critical role to play in enabling nations to plan for and monitor progress towards the SDGs.

92. There is a subset of the SDG targets where the application of Earth observation and geospatial data and information is key to accurate, consistent and relevant measurement of progress. For example:

- *Goal 6* is to ensure the availability and sustainable management of water and sanitation for all. Several of the proposed indicators could use Earth observation and geospatial information as their data source. For example, proposed indicator 6.4 for water efficiency and proposed indicator 6.6 on water related ecosystems.

- *Goal 9* is to build resilient infrastructure, promote inclusive and sustainable industrialization and foster innovation. Proposed indicators 9.1.1 (Share of the rural population who live within 2km of an all season road) and 9.c.1 (Percentage of the population covered by a mobile network, by technology) each requires geospatial information about the location of the population in conjunction with the location of physical or technological infrastructure.

- *Goal 14* is to conserve and sustainable use the oceans, seas and marine resources for sustainable development. Proposed indicators 14.2 (marine and coastal ecosystems) and 14.3 (ocean acidification) could both make extensive use of Earth observation and geospatial information as their data source.

- *Goal 15*, to protect, restore and promote sustainable use of terrestrial ecosystems, sustainable manage forests, combat desertification and halt and reverse land degradation and halt biodiversity loss, is where Earth observation and geospatial information could be a particularly valuable data source.

93. Earth observation data already underpins global efforts to monitor climate change, report on deforestation and monitor agricultural productivity.

94. Earth observation data, and information derived from it, have a key role to play in enabling nations to effectively monitor progress towards the SDGs as they provide a truly objective view of the Earth and how it is changing. It is important that indicators can be consistently tracked over long periods at varying scales, and that comparability across nations can be achieved.

95. All Earth observation data can play an important role in tracking SDG progress. However, satellite Earth observation data is increasingly available globally on a free and open data policy, particularly at moderate resolution. This gives it significant potential to support sustainable long-term monitoring of relevant SDGs.

96. The intergovernmental Group on Earth Observations (GEO)[8] brings together over 90 nations to exploit the potential of Earth observation data to deliver societal benefits. GEO has been in operation for ten years, and was recently extended for a further ten years.

---

[7] A/RES/70/1
[8] http://www.earthobservations.org/

A key focus of the next decade of GEO will be enhancing partnerships with UN institutions and frameworks to ensure new services transition into 'operational' use on a global basis, building capacity in Africa, Asia and South America, and ensuring Earth observations are applied to the SDGs.

97. The use of geospatial data, including geospatially enabled statistics and information about the built environment, will also have an important role to play in compilation of the SDG indicators.

98. At its fifth session held in New York from 5-7 August 2015, the UN Committee of Experts on Global Geospatial Information Management (UN-GGIM) noted the critical role of geospatial information and Earth observation data in providing: "objective, comprehensive and authoritative data and information [that] will be needed in a timely manner to support sustainable development policy and decision making at all levels;"

99. The initial proposed SDG indicators and stakeholder comments indicate that the role of Earth observations and geospatial data and information, or the integration of these official data sources with more traditional statistical data sources such as Census, survey and administrative data, was not adequately considered. A significant number of the proposed indicators did not reflect current and emerging practice and technology and therefore would not fully harness the richness and depth of this data.

100. Further to this, the current indicator specifications do not capitalise on the significant opportunities Earth observation and geospatial data and information, particularly when integrated with traditional statistical data sources, provide. Australia has recommended that action be taken to address this or the SDG indicator and reporting framework will not be as efficient and effective as it could be.

101. As well as making it possible for Earth observations to be applied in more sophisticated ways to new challenges, new technologies like the Australian Geoscience Data Cube[9] are significantly lowering the technical barriers to using Earth observation data alongside more traditional statistics.

102. Data within the Australian Geoscience Data Cube is compatible with the emerging Discrete Global Grid System[10] standard under development within the Open Geospatial Consortium. This Grid System could also be used to manage and integrate other data sources, not just Earth observation data. Geocoding data is one of the five dimensions of the global statistical geospatial framework being developed by the Expert Group on the Integration of Statistics and Geospatial Information. Geocoding traditional statistical data sources such as census, survey and administrative data, to an implementation of the Discrete Global Grid System is an option for addressing the technical challenges of integrating statistical and geospatial data in a flexible and integrated way.

103. In response to 'big data' and modernisation opportunities, many NSIs are putting considerable effort to develop capabilities in managing, processing and analysing geospatial data and information, and increasingly this includes Earth observation data. This is often being done in parallel with their work exploring other data sources, such as administrative data and other 'big data' such as telecommunications data. This effort has required NSIs to look at the modernisation of their infrastructure and policies, as well as developing new skill sets beyond the traditional areas associated with interrogating and analysing standard census/survey data.

104. That said, further consideration needs to be given to the mechanisms for ensuring access to this Earth observation and geospatial data and information at both global and national scales. Similarly, consideration needs to be given to the development of technical

[9] http://www.datacube.org.au/
[10] http://www.opengeospatial.org/projects/groups/dggsswg

capabilities to access and analyse this information. Addressing these access and capability issues are particularly important from the perspective of developing countries so that they can effectively contribute to and benefit from Earth observation and geospatial data and information with the context of the SDGs and their own development objectives.

105. The United Nations Statistics Division Global Working Group on Big Data for Official Statistics[11] aims to meet the expectation of the society for enhanced products and improved and more efficient ways of working, and to support the monitoring of the SDGs by improving timeliness and relevance of indicators without compromising their impartiality and methodological soundness. The Global Working Group has eight task teams, two of which are relevant here.

106. The United Nations Statistics Division Global working Group on Big data for Official Statistics Task Team on Satellite Imagery is currently applying their technical expertise and resources to make use of satellite imagery data for official statistics; and the Task Team on Linking Big Data and Sustainable Development Goals is considering the use of Earth observation data in the SDG indicator process. To be successful, both must engage with experts in the geospatial community to maximise the full benefits of Earth observation and geospatial information and technology.

I. Output Geographies

107. The dissemination of statistics requires an increasing number of output geographies that can match the wide number of user needs for statistics as well as the needs to balance quality, comparability, consistency and disclosure.

108. No single approach to statistical dissemination exists and where attempts have been made towards identifying the best approach to disseminating statistics such as the work of the UN-GGIM Expert Group to compare the benefits of producing statistics on grids and administrative geographies have found that no single approach can meet the needs of statistical users.

1. Grids

109. Statistical grids are a system of statistical output areas composed of equal sized, normally squared grid cells with varying grid cell sizes. In a cross-border context 1km² grid cells have become a quasi-standard and represent a good compromise between the demand for detailed and flexible statistics and data protection concerns. Most commonly statistical grids are used to report population counts or population densities but in principle are suitable for the dissemination of any type of spatially referenced statistics.

110. For population and housing census purposes, a population grid is defined as a geo-referencing framework for population in the form of a grid net with fixed and unambiguously defined locations of equal-area grid cells[12]. In this format they have been recognised by various statistical communities as a useful output system including the European Statistical System[13], the European Forum for Geography and Statistics EFGS[14], the UNECE in their census recommendations for the 2021 census.

111. It is important to note that in their strict sense grids are just a geographical output system used to aggregate existing point based data. As such they represent just one way of

[11] http://unstats.un.org/unsd/bigdata/

[12] http://www.unece.org/fileadmin/DAM/stats/publications/2015/ECE_CES_41_WEB.pdf

[13] https://circabc.europa.eu/d/a/workspace/SpacesStore/fd349927-3c7d-435a-8b80-4ace48daf646/D_GIS_105%20GISCO-TF-Report-V_3.doc

[14] http://www.efgs.info/

presenting statistical data and do not aim to replace other output geographies. This was confirmed by the UN EG-ISGI[15]. In the absence of a point based statistical data as a source data for statistical grids, grids might be produced from disaggregated lager statistical geographies using auxiliary information.

112. The approach to statistical output areas used for statistics varies a lot between Member States. While in some countries the output system has remained stable for decades, other countries introduce frequent changes into their system to meet demands from politics, e.g. for efficiency gains in public administration or to reflect demographic trends. These changes make it necessary for output systems to be revised on a regular basis. This instability of the territorial building blocks for statistics makes the production of time series difficult and their use difficult for many types of statistics and as such limits its use for analysis or spatial planning.

113. Also, the extreme difference in area and population of territorial units across countries make any type of spatial analysis and comparison between countries difficult, as e.g. vast but almost empty areas in the north of Scandinavia are compared with small but densely populated megacities.

114. Hence, since several years, spatial analysts in NSIs, researchers and spatial planners have been suggesting to complement the current territorial system for official statistics based on administrative areas with a neutral, hierarchical and stable system. Output areas based on equal population counts represent a step in the right direction but do not overcome some of the intrinsic shortcomings of statistics based on administrative units such as instability over time or the variety in area size. Hence their focus on population numbers may still limit their usage for analytical purposes and statistical products.

115. To respond to this criticism of territorial frameworks based on administrative units or other irregular tessellation systems, grid statistics have been proposed as an alternative output system. It is important to note though, that grids represent just one way of presenting statistical data and do not aim to replace other output geographies. This was confirmed by the UN EG-ISGI[16].

116. Statistical grids are a hierarchical, neutral and stable system composed of equal sized grid cells. Statistical grid data help to more accurately capture the spatial extent and magnitude of phenomena. It is easy to construct a hierarchical system of grids, where the size of grid cell is in proportion to the size of the study area and the resolution of the phenomenon, ranging from local to global. The transition from point based microdata to grid data becomes a continuum.

117. Socioeconomic statistics at grid level are easily linked with scientific data and data from spatial modelling, which are often created based on grid cells or other finite elements, e.g. meteorological data, biodiversity information, or data from climate modelling. This is particularly important for some of the most pressing challenges for sustainable development such as mitigating climate change risks, or protecting our natural capital or preventing hazards.

118. Starting in the Scandinavian countries in the 1970s several statistical offices have added grid statistics to their product portfolio, mainly in the area of population statistics. To report population counts or population densities are still the most common use of statistical grids but in principle they are suitable for the dissemination of any type of spatially referenced statistics.

119. Population grids have now and gained wide recognition in the EU and beyond as a complementary system for geo-referencing and disseminating statistics in particular for

[15] http://ggim.un.org/docs/meetings/2nd%20UN-EG-ISGI/EG-ISGI-Second%20Meeting-Summary-.pdf

[16] http://ggim.un.org/docs/meetings/2nd%20UN-EG-ISGI/EG-ISGI-Second%20Meeting-Summary-.pdf

population and housing censuses[17], including the European Statistical System[18], the European Forum for Geography and Statistics EFGS[19], and UNECE in their census recommendations for the 2020 census round.

120. Grid cells, provided they have a suitable size, can be joined together to create virtually any functional output area for statistics based on objective criteria, e.g. to statistics for mountain areas, school districts, coastal areas, river basin districts, etc. This is usually impossible with other territorial systems and makes the grid system extremely valuable for analysis and spatial planning.

121. In a cross-border context 1km² population grids have become a quasi-standard and represent a good compromise between the demand for detailed and flexible statistics and data protection concerns.

122. It is important to note that grid statistics are easy to produce from geocoded microdata from full enumerations like in censuses, or from registers or other administrative data sources containing virtually all entities. However achieving statistical representativeness for small areas from survey data is extremely costly. This means that for surveys, grid cell sizes would have to be enormous and reach 10s or even 100s of km, thus devaluing the grid advantages.

123. Big Data may have the potential to overcome these limitations, and may help to reach the high spatial resolution necessary for grid statistics. In the absence of a point based statistical data as a source data for statistical grids, grids might be produced from disaggregated lager statistical geographies using auxiliary information.

124. On the dissemination side the correct handling of disclosure control in any small area including grids quickly becomes an issue. This is very relevant in population statistics and census information where crossing just a few population characteristics in a small area with low population quickly leads to the risk of disclosure of individuals. Within the global community a discussion is taking place on the sensitivity of various population characteristics and on suitable disclosure control thresholds. For the EU a specific action will be launched in 2016 on statistical disclosure control measures. This action will also address the specific challenges arising from introducing grid statistics.

2. The GEOSTAT 1 project on a pan-European population grid

125. The latest housing and population census in 2011 has been a catalyst for this trend, as many NSOs used point based spatial reference frameworks, in most cases address and building registers during the preparation and execution of the Census. Preserving the location element with the statistical information allowed NSOs to aggregate data to any type of area including grid cells.

126. To expand and accelerate this trend to grid statistics, Eurostat launched the GEOSTAT initiative in cooperation with the European Forum for Geography and Statistics (EFGS). The goal of the first project under GEOSTAT, GEOSTAT 1, was to establish a methodology[20] for a pan-European 1km² population grid using the most detailed existing census information from the 2011 population census. Based on this methodology NSO developed national grid data sets. Ideally the data were georeferenced to a point at individual record level, or at least aggregated or disaggregated.

---

[17] http://www.unece.org/fileadmin/DAM/stats/publications/2015/ECE_CES_41_WEB.pdf

[18] https://circabc.europa.eu/d/a/workspace/SpacesStore/fd349927-3c7d-435a-8b80-4ace48daf646/D_GIS_105%20GISCO-TF-Report-V_3.doc

[19] http://www.efgs.info/

[20] http://www.efgs.info/geostat/1B/frontpage/final-technical-report

127. In addition, GEOSTAT 1 has for the first time developed a specific quality documentation[21] dealing with the specifics of grid statistics on the interface of statistics and geography. The final European GEOSTAT 2011 grid dataset was compiled from 28 national grid datasets based on national census information, and gaps were filled with modelled information for in total only four countries representing less than 2 million citizens or less than 1% of the population of the EU.

128. The demand for these grid data has been high and the data have been used already in the definitions of territorial classifications such as the Degree of Urbanisation. Using a population grid helped to take account of the actual spatial distribution of population in LAU2 and NUTS areas and helped to avoid the dilution effect of average density in large area administrative units with a highly concentrated population.

129. Other countries outside the EU have adopted the GEOSTAT methodology and have created prototype national population grids, e.g. Australia.

3. Administrative Geographies

130. The majority of statistical data continue to be needed at the administrative level to support the delivery of services. This means that despite the advantages of using grids and other statistical geographies for disseminating data, administrative geographies will continue to remain a core geography for statistical production.

131. Producing a harmonised approach to administrative geography is difficult because they serve a diverse range of functions and have been built around a diverse set of criteria. This makes it difficult to analyse statistical data for administrative areas at the European or global levels.

132. To attempt to resolve the problem Eurostat set up the NUTS classification as a single, coherent system for dividing up the EU's territory in order to produce regional statistics for the administrative geographies. Having originally been implemented under an agreement, the NUTS boundaries were given legal status through the European Commission Regulation (EC) No 1059/2003. There are three NUTS levels consisting of 98, 276 and 1342 regions respectively and these cannot be updated for at least three years to give some stability for time-series analysis.

4. Hybrid Geographies

133. The instability of administrative geographies for time-series analysis and the inconsistency of grid populations for comparison of populations means that NSIs are increasingly looking towards other geospatial alternatives for publishing official statistics.

134. Most commonly this seems to consist of a hybrid approach where grids or other small building blocks such as postcodes can be aggregated on a set of criteria designed to overcome the limitations of grids and administrative geographies.

135. Within the UK the two most notable examples are the Built-Up-Areas – a statistical geography built from aggregations of 500m cells based on the relationship between the built up environment within each grid cell, and the Output Areas (OAs) – the smallest geographic unit for which non-disclosive statistics can be produced. This is because each OA is built around a minimum threshold of 100 resident population and 40 households. There is also a loose attempt to create social homogeneity within each OA using tenure and accommodation type to group postcodes within similar characteristics together in the OA design.

[21] http://ec.europa.eu/eurostat/web/gisco/geodata/reference-data/population-distribution-demography

136. It is likely that this sort of hybrid approach to statistical geography will become more prevalent as NSIs try to balance the need to release increasing numbers of statistics with the need to prevent disclosure by differencing on a wide range of geographic bases.

J. Standards

137. To ensure there is international comparability and coherence, standards for what is meant by ‘location’ will be required. In some cases, this location could be a region such as a city, a province or a jurisdiction. In some cases, location could be a classification such as urban, rural, remote and sparsely populated.

138. There is currently no globally agreed standard or classification for many of these possibilities. Furthermore, whilst there may be agreement within domains that will support certain SDGs (such as economic performance indicators proposed for Goal 8 Promote sustained, inclusive and sustainable economic growth, full and productive employment and decent work for all), agreement across the domains for which all the SDGs span (such as economic performance, societal progress and environmental sustainability) adds further complexity.

139. For instance, whilst the OECD through its Territorial Development Policy Committee and Working Parties has developed a definition of Functional Urban Areas[22], this has been developed by the OECD and the European Union and does not apply globally. Similarly, the OECD has made use of the UNECE land classification[23] designed for environmental indicator reporting, but other indicator systems make use of alternative classifications such as the System of National Accounts which define land according to the service it is providing.

140. The statistical community has a long and well-established mechanism for developing and endorsing globally-agreed statistical standards, classifications and frameworks e.g. standards for internationally-comparable price statistics and GDP. The geospatial community also has mechanisms for developing and endorsing standards, particularly through standards groups such as the ISO, the Open Geospatial Consortium and W3C. These technical communities will need to work together if international comparability and harmonisation across the seventeen SDGs is desired.

141. Interoperability and comparability within and between geospatial data and information and statistical data and information, using a globally endorsed framework, will be key to a robust and repeatable set of globally-accepted indicators. The United Nations Statistical Commission and UN Committee of Experts for Global Geospatial Information Management have established an Expert Group on the Integration of Statistical and Geospatial Information that is tasked with developing a global statistical geospatial framework. This framework is modelled on the Australian Bureau of Statistics Statistical Spatial Framework[24] and aims to assist countries to organise, integrate and apply geospatial enabled statistical data and information. A key element of the framework is metadata interoperability which, at its core, is about the integration of standards between the statistical and geospatial communities.

[22] http://www.oecd.org/gov/regional-policy/Definition-of-Functional-Urban-Areas-for-the-OECD-metropolitan-database.pdf

[23] https://stats.oecd.org/glossary/detail.asp?ID=1502

[24] http://nss.gov.au/nss/home.nsf/pages/Statistical%20Spatial%20Framework%20Homepage

K. Dissemination – INSPIRE

142. Most of the activity that has taken place on the dissemination of geospatial statistics has come from the European Commission's INfrastructure for SPatial Information in the euRopEan community (INSPIRE) legislation. INSPIRE was a regulation introduced by the Commission in 2007 to standardise the approach to the dissemination of geospatial data within Member States to support cross-border response to environmental issues.

143. INSPIRE sets out core data themes needed to respond to such an issue and then sets out how those data themes can (and should) be delivered in an interoperable way.

144. For the statistical community this was particularly represented by the Statistical Units and Population Distribution – Demography themes that provided the output geographies for the dissemination of statistics and the population data that could supplied for those output geographies.

145. The problem for INSPIRE is that it is a geospatial initiative built around the availability of Web Mapping Services (WMS) and Web Feature Services (WFS) to make the data available. This works well for the dissemination of statistical units (that are geographic boundaries) but does not work for the dissemination of the population distribution – demography theme. It is only possible to disseminate this theme through the integration of these datasets with geospatial data.

146. There has been some developments within this area to investigate potential methods for disseminating integrated statistics and geospatial information. Most notably the use of the OGC Table Joining Service (TJS) within the Netherlands[25] and linked data within the UK[26] (further details of this are within section 19 of this review).

147. Dissemination remains a problem for geospatial statistics due to the lack of standard interoperability (as set out in section 17 of this review), and this is something that will need further work by the statistical and geospatial communities if the vision of integrated statistical and geospatial data is to be achieved.

148. The other problem for INSPIRE was that harmonised statistical datasets already exist at the European level through initiatives such as GEOSTAT and any data specification that INSPIRE could provide for the statistical data theme could never achieve the level of detail and interoperability that already exists. At this point, INSPIRE risks becoming a burden by asking statistical institutes to duplicate population data and supply data for a less detailed INSPIRE standard alongside the existing more detailed dissemination. This is the risk of initiatives around the integration of geospatial information and statistics diverging rather than converging.

L. Dissemination – linked data

149. Also within the GSBPM phase of dissemination is the need to consider the formats through which geospatial statistics and geospatial information. In the context of the CSPA, catalogues are seen as key enablers. They provide lists and descriptions of standardised artefacts or code lists, as well as information on how to obtain and use them.

150. There are many ways of providing catalogues of managed code lists and their descriptions and many of these are code lists are well established and used at the international level. Within the UK the catalogues of geographic codes for statistics are

[25] http://www.geonovum.nl/onderwerpen/services/documenten/rapport-joining-tabular-and-geographic-data-merits-and-possibilities-table-joining

[26] http://statistics.data.gov.uk

being provided by linked data and this format is increasingly being implemented for dissemination because it provides the opportunity to define the relationships between managed code lists and disparate data sources making them more efficient to query and exploit.

151. As a data format, linked data takes an existing code list and publishes it online in an open and accessible format that can be interrogated in the same way as a catalogue. Placing the data online makes it easier to find and access and the existing linked data codes are already used to provide quality assurance of statistical tables.

152. The World Wide Web Consortium (W3C)[27] has established a 'Spatial Data on The Web' Working Group to investigate how geospatial information can be published online more effectively and how it can be linked to other data (such as statistics) through formats such as linked data. The Working Group is aiming to deliver best practice guidance and ontological models for linking geospatial data to statistics (amongst other data) by the end of 2016.

V. Country practices

A. Australia

1. Sustainable Development

153. In September 2015, the Australian Bureau of Statistics (ABS) coordinated an Australian Government response to the Inter-Agency and Expert Group on the Sustainable Development Goal Indicators (IEAG-SDGs), established by the United Nations Statistical Commission to propose indicators to monitor the seventeen goals through approximately 169 targets and at least that number again of indicators.

154. This process highlighted several key issues regarding the production of indicators using geospatial data including Earth observation and other geospatial information. The ABS received strong input on this theme from Geoscience Australia (GA) and Commonwealth Scientific and Industrial Research Organisation (CSIRO) in particular.

155. The significant technical expertise and resources that were once barriers to use of Earth observations are now much lower. This is important in developing countries as it will enable more effort to be put into analysing and applying data rather than collecting and organising it. This technical expertise and resource can also be beneficial to the monitoring and reporting of SDGs, particularly in the context of Earth observation data being a 'big data' source.

156. For example, the Australian Bureau of Statistics has used its 'statistical' technical expertise to develop a method for estimating crop yields from Earth observation satellite data in the Australian Geoscience Data Cube[28]. However, further exploration of this method using 'geospatial' technical expertise is needed to ensure that both technical expertise and data, regardless of their source, are part of the 'Big Data for Official Statistics' journey.

[27] More detail is available at http://www.w3.org/2015/spatial/wiki/Main_Page

[28] http://unstats.un.org/unsd/trade/events/2014/Beijing/presentations/day2/morning/1.%20Satellite%20Imagery%20talk--Siu-Ming%20Tam.pdf

2. Disclosure Control

157. The Australian National Statistical Service (NSS)[29] has a Confidentiality Information Series that provides information on the obligations and identification risks relating to data releases, as well as methods that can be used to confidentialise data. The Confidentiality Information Series can be found on the NSS website. A more detailed paper on the ABS website – Research Paper: A Review of Confidentiality Protections for Statistical Tables, Jun 2005 (ABS Cat. No. 1352.0.55.072) – examines some of the confidentialisation techniques that can be used to protect privacy. The remainder of this paper focuses on the specific privacy risks for data with a location or region component.

158. The Statistical Spatial Framework (SSF) recommends that unit records in socio-economic datasets be geocoded with location coordinates (i.e. latitude and longitude) and an Australian Statistical Geography Standard (ASGS) Mesh Block[30] code. This geocode information is usually obtained through geocoding the location address information for each statistical unit in the dataset. While de-identification of a dataset will generally remove the address information, the coordinate information and possibly the Mesh Block code could be combined with other information in the dataset to identify an individual or organisation. Due to the precision of coordinate information, it is generally recommended that it not be released in combination with other characteristic information contained in socio-economic datasets.

159. In specific instances, it may be necessary or desirable to release data at the unit record level. In these instances the data should be treated as 'microdata' and the specific guidance around the release of this type of data should be carefully considered. For further guidance see the Confidentiality Information Series Information Sheet 5: 'Managing the risk of disclosure in the release of microdata'.

160. The National Address Management Framework (NAMF)[31] includes provision for interchange of address and coordinate information between organisations where other information is not included. This is documented in the NAMF 'Address data interchange standard', which can be found on the NAMF webpage.

161. The Statistical Spatial Framework (SSF) recommends that regional data released or made available from any socio-economic dataset should include Australian Statistical Geography Standard (ASGS)[32] regions. ASGS regions are the common geography in the SSF. The SSF acknowledges that releases of data from these datasets may also include additional regional breakdowns that are not part of the ASGS (e.g. school catchments or Medicare local regions).

162. It is recommended that a careful assessment should be made before releasing information for ASGS Mesh Blocks and any other very small area or population regions. While the Mesh Blocks may be available through the address coding process, there are usually only a limited number of persons or organisations in each Mesh Block. These low numbers mean there is a high likelihood that individuals or organisations would be able to be identified through the use of only a limited number of characteristics, creating substantial risk of a breach of privacy through disclosure of private information. Similar concerns apply for other very small area or population regions. If data is released for Mesh Blocks, the information provided in the remainder of this paper is particularly relevant and should be carefully considered.

[29] More information about the National Statistical Service is available on the NSS website.
[30] More information on Mesh Blocks is available on the Australian Statistical Geography Standard (ASGS) webpage.
[31] More information is available on the National Address Management Framework (NAMF) webpage.
[32] More information is available on the Australian Statistical Geography Standard (ASGS) webpage.

163. It is generally recommended that data only be released where it is aggregated (i.e. summed or grouped together) for medium to large geographic regions. By aggregating data into regions the private information for an individual or organisation is combined with other private information; as more data is combined together it becomes increasingly difficult to identify the separate pieces of private information. The process of aggregating data by regions or by other classification groups (such as industry classification) is an important and fundamental confidentialisation tool that is used to protect privacy.

164. The region type selected for releasing data should have a large enough number of statistical units (e.g. persons or organisations) in each region to maintain confidentiality for the majority of the variables for which data is planned to be released. The size of the region must also be balanced against the regional analysis needs of the users of the data. Data confidentialisation techniques can then be used to manage the remaining instances where privacy risks exist.

165. Selecting a region type usually requires testing possible options to determine if any particular region type will contain large amounts of statistical information that requires further confidentialisation. The information provided below explains how to identify and manage these privacy risks.

166. Releasing data using geographic regions helps to manage privacy risks but does not completely resolve these issues. Geographic regions are only one set of variables in a dataset and the other characteristic variables can be used in combination with each other, or in combination with region, to identify an individual or an organisation. Therefore, each possible cell of data needs to be tested to assess the privacy risks.

167. A range of methods are available to identify and manage privacy risks when preparing data for release. These methods can be applied equally to data that is broken down by region or by other characteristic variables, or a combination of these.

168. The "Confidentiality Information Series Information Sheet 4 - How to confidentialise data: the basic principles" provides information on these methods. When applying these confidentialisation methods there are a few specific issues that need to be considered when the data to be released is broken down by region. These are discussed in the next section.

169. Other confidentialisation methods for managing privacy risks when releasing data in a geospatial format have been used by a range of organisations. These methods include:

- modelled characteristic information, which removes the risk of an individuals or organisations actual information being released, and
- data generalised to a grid system, including smoothing of data across a grid to remove extreme or unique values.

170. There are also privacy risks that are specific to regionalised data released from socio-economic datasets. It presumes readers are somewhat familiar with the privacy risks and confidentialisation methods outlined in the "Confidentiality Information Series Information Sheet 4 - How to confidentialise data: the basic principles".

171. Methods that can be used to ensure the confidentiality of the data where overlap areas create privacy risks include:

- modifying the regions to be released to eliminate the areas of overlap,
- combining regions to eliminate the areas of overlap, and
- using techniques, such as category collapsing, suppression or perturbation which are described in the "Confidentiality Information Series Information Sheet 4" (noting the issues associated with suppression outlined below)

172. A simple method of removing identified privacy risks is to combine or merge one or more sub-regions where privacy risks exist. This allows all of the data available to be released, but obscures the detail for the combined sub-regions. A related method to resolving identified privacy risks is to modify the boundary that data is being released for, to eliminate the area of overlap containing records. Both of these methods require an effect way of identifying where geographic differencing risks exist to ensure all issues are addressed.

173. An alternative to modifying the geography (through merging or boundary adjustment) is to suppress the data within regions where there is a risk of disclosure. The need to have sub-regions supressed within any larger region applies to any future release of data for regions built from these sub-regions. This may mean restricting the subsequent release of data for larger regions that contain only one of the suppressed sub-regions.

174. An alternative approach to this problem is to combine or merge the sub-region that has the privacy risk with one of the other sub-regions (as described above). This method allows all of the data available to be released, while also obscuring the detail for the sub-region that contains the privacy risk.

175. The complexity of these geographic differencing issues highlights the value in using nested, hierarchical regions, such as those included in the ASGS. Regional hierarchies of this type help to limit and manage this complexity.

176. Perturbation is another very effective method that protects against geographic differencing by randomly introducing small changes to cell counts or values. If differencing occurs, the degree of confidence of an exact match is substantially reduced. Perturbation is effective at all levels of differencing even when the difference is one unit. The Australian Bureau of Statistics uses perturbation for statistical releases and considers the use of perturbation to be a suitable technique for eliminating the risk of releasing identifiable statistics while maximising information that can be released.[33, 34]

177. Using multiple region types (that are not nested, hierarchical regions) when releasing data can create a very large number of privacy risks to assess and manage. This is particularly relevant when location coordinate information is used to allocate unit records to the regions being released. Using location coordinates creates a high degree of certainty of identification if only a few statistical units (e.g. household or businesses) are geographically differenced within any of the overlapping areas.

178. To illustrate, consider a hypothetical example of releasing data on two region types – School Catchment Zones and ASGS Statistical Areas Level 2. As each of these boundaries is designed for a different purpose they will cover different areas, which will result in overlaps between regions. Any of these overlaps may pose a privacy risk due to the potential for geographic differencing. In some instances, these overlaps may be relatively small. Where the area covered by these overlaps is small, the overlap area or areas may only include a very limited number of individuals who may be easily identifiable from the resulting geographically differenced data (other similar examples apply to data for organisations).

179. As with the simpler examples discussed above, the region being released must be compared with the regions included in previous releases to identify each instance of potential geographic differencing created by overlapping regions. This assessment may

[33] Leaver, V. (2009) "Implementing a method for automatically protecting user-defined Census tables", Joint UNECE/Eurostat work session on statistical data confidentiality, Bilbao, Spain.

[34] Fraser, B. and Wooton, J. (2005) "A proposed method for confidentialising tabular output to protect against differencing", Joint UNECE Eurostat work session on statistical data confidentiality at Geneva, Switzerland, 9-11 November 2005

include a large number of overlap areas and may include complex cases where the boundaries of a number of individual regions, across several region types, create many sets of overlapping areas.

180. Identifying where regions overlap with each other is a relatively simple task using geographic information system (GIS) software. Identifying the level of risk that each one of these overlap areas represents is more challenging, the differenced data for each overlap area must be calculated and assessed for privacy risks.

181. Methods that can be used to ensure the confidentiality of the data where overlap areas create privacy risks are the same as the Methods to ensure confidentiality included in the section above.

B. United Kingdom

1. Addressing

182. ONS have begun work on the address register (AR) for the 2021 Census. The 2021 Census AR will contain all addresses (e.g. residential, communal establishments, businesses) in a single register. The aim of the 2021 AR is go beyond providing a list of addresses and classifications; instead the ambition is to collate a range of address intelligence.

183. Address intelligence refers to information on the occupancy, use and characteristics of each address. Examples of address intelligence include; supply of broadband, rental/ownership data, rate of occupant churn, barriers to entry, second home/holiday home use.

184. ONS believe that address intelligence can have a wide range of applications; improve field enumeration, enable follow-up prioritisation, facilitate edit and imputation methods, facilitate estimation.

185. The 2021 AR will be a legacy product and will become the ONS corporate AR.

186. One of the challenges when collating a range of address data is the linking accurate linking of records. ONS are working with colleagues across government to develop open source code for address matching.

187. Address matching is not only key in the development of an enhanced AR, it is also vital when linking person records. Address identifiers and geographical identifiers (e.g. Postcode) are used when linking person records across administrative data. It is critical that these fields are accurately referenced.

188. ONS are leading work to standardise the formatting and referencing of addresses across Government. Government departments hold an array of datasets; a common attribute on many of these datasets are addresses. ONS are working to create a central address services for the maintenance and referencing of addresses across government; this will enable all government data to be quickly and efficiently combined at address level.

189. This work is being conducted in collaboration with Government Digital Services (GDS). As such, the ambition is to help create a single reference point for the use and specification of addresses across all government activities (e.g. vehicle registration, passport applications, tax and benefit records).

190. ONS Methodology are working to highlight the value of address data and intelligence in the production of administrative data based population statistics. ONS Methodology believe that knowledge of small area address characteristics (e.g. number of residential addresses, type and size of addresses, signs of occupation data (e.g. energy

usage), rental rates, occupant churn) can be used to understand and quality assure person records found on administrative data.

191. Knowledge of address and area characteristics has the potential to provide a means of assessing the quality and coverage of administrative data.

2. Big Data

192. ONS has made a particular effort to explore the potential of mobile phone data to gain more timely and granular information about movement of people. There are a number of barriers to access, an important one being legitimate concerns around privacy and public perceptions. The general approach to mitigating these concerns is through aggregating data using common spatial units. Thus, geography also has a key role to play in helping to protect privacy although this raises other issues, notably a lack of transparency in exactly how data have been compiled.

193. Another area ONS is starting to explore is the use of satellite imagery. This involves using these data to systematically identify certain types of hard to enumerate addresses in preparation for the 2021 Census. The current focus is on identifying caravan parks, which typically have unusual occupancy patterns. However, there are many other applications of satellite data for policy-making including statistics about land use, the impacts of climate change and other human impacts on the environment.

194. The availability of more administrative and big data is expected to have far-reaching implications for the future of the Census. Although ONS is committed to a full census in 2021, future census-type outputs could be produced using a completely different approach. Depending on feasibility, these could be produced on a more frequent basis, combining a range of administrative, survey and big data sources. This in turn could have implications for the production of the statistical units on which Census outputs are based.

195. Census Output Areas and related geographies are reviewed as part of the decennial Census cycle. They are updated from new Census data using design criteria such as, population size thresholds, regular shape and co-linearity with natural boundaries while also minimising change. Thus, these units of measurement are to a considerable degree influenced by the population changes they are designed to measure. If these census type outputs are produced with greater frequency, then some careful thought needs to be given as to how these statistical units will be maintained in future.

196. In summary, geography is very relevant to big data because so much big data has a spatial dimension. Also, in the context of official statistics, geography is about structure while a lot of big data is essentially data that lacks structure. Thus, geography is an important means of making sense of data in the emerging age of big data.

VI. Issues and challenges

A. Consolidation of activity

197. The largest challenge facing the statistical community is how to consolidate a diverse set of activities into a clear, coordinated approach to the creation of geospatial services for official statistics. There is a natural hierarchy to the activity that is currently taking place. The development of the Global Statistical Spatial Framework provides a strategic layer to the work and the integration of geospatial services into the GSBPM provides the technical framework to implement that strategy.

198. Despite this, there is still a risk, given the large number of stakeholders, that work on the integration of statistics and geospatial information begins to diverge, with each activity

having its own objectives that are similar to – but do not align with – other activities taking place within the same community.

199. In particular, activities such as the UN-GGIM:Europe implementation of the core data working package including no statistical representation to ensure core geography for statistics was captured and the lack of consideration for how UN-GGIM and GEOSTAT can align with the CSPA (and how CSPA can align with GEOSTAT and UN-GGIM).

200. There is also now a proposal to include work packages on the European Commission's ISA2 (Interoperability Solutions for European Public Administrations) Work Programme on the following topics:

- Sharing Statistical Production and Dissemination Services and Solutions in the European Statistical System
- European Location Interoperability Solutions for E-Government

201. Whilst the recognition of the importance of the integration of statistics and geospatial information is welcomed, the overlaps between these proposals and existing activities mean there is a risk of duplication which with limited resources is something that the international statistical community should be resisting as strongly as possible.

202. It is critical that the international statistical community recognises the existing work being done through UNECE, UN-GGIM, UN-GGIM:Europe, the Open Geospatial Consortium, Eurostat, GEOSTAT and the European Forum for Geography and Statistics amongst others, and works with all of these stakeholders towards a common approach to the integration of statistics and geospatial information.

203. There should be some consideration of how oversight can be maintained of all these activities and whether the UNECE could be in a position to provide this role. Without the willingness of an international statistical body to support the consolidation of these activities into a single approach to geospatial statistics, a risk remains that these activities will start to diverge towards discreet silos of work.

B. Ambiguity of terminology

204. The UN EG-ISGI in its first meeting in November 2013 concluded that[35] "… in order to meet the objective of providing a forum for coordination and dialogue among representatives of both statistical and geospatial communities, with a view to developing and advancing the implementation of a global statistical-geospatial framework, it was important to first share knowledge of existing terminologies and practices."

205. A precise unambiguous definition of geospatial methodologies and practices relevant for statistics is one of the conditions for semantic interoperability between statistical and geospatial information and also relevant all along the statistical production process, from the design to the dissemination stage. Producers of statistics, users of statistics and providers of auxiliary data including geospatial data need to have a common understanding of how statistical data and geospatial data have been produced, what is their quality, how they were integrated and for which geographies they have been produced. Accepted common definitions are also vital for a good cooperation and communication between statisticians and geospatial experts.

206. Currently common terms might be ambiguous depending on the context. As an example, while "Statistical unit" in statistics means an object of observation such as a

---

35 http://ggim.un.org/docs/meetings/UNSG_EG/ESA_STAT_AC.279_L4_Report%20of%20the%20EG%20Meeting_rev.pdf

person or a business, INSPIRE defines "statistical unit" as a geographical area used for statistical reporting or collection of statistics.

207. Currently there is no agreed international definition of geospatial concepts and methodologies which are relevant for statistics. In the second meeting of the EG-ISGI in May 2015, Eurostat presented a first draft[36] of definitions and synonyms based on a screening of national, UN and European papers, see[37] for the list of sources. The Eurostat paper identified the following main areas that would require an agreed definition of terms, concepts and practices:

- Definition of the result of the integration of statistical and geospatial information, the actual statistical product.
- Definition of geospatial methods relevant for the production of statistics such as integration, linking, geocoding, georeferencing, disaggregation and aggregation.
- Definition of geographies relevant for the dissemination and collection of statistics, such as statistical output geographies, point based geographies, functional geographies, grid based geographies.

208. Some of the concepts such as statistical geographies are also relevant for the definition of core geospatial information (see other chapter e.g. on core data) and should be aligned. Thus an agreed set of definitions is also relevant for the future statistical geospatial framework.

209. The UN EG-ISGI welcomed the general approach set out in the paper and agreed to review the proposal in a small editorial board of experts from both the statistical and geospatial communities. After agreement by the UN EG-ISGI the terminology repository should eventually be hosted in a website of UN-GGIM.

C. Disclosure Control

210. Confidentiality remains a challenge to the statistical community and the increased use of alternative data sources (such as administrative and big data linking) will only add to the problems of protecting confidentiality within data.

211. Existing statistical methodologies for disclosure control will need to be reconsidered in the context of the increasing use of spatial data to ensure that location does not threaten the ability to identify attributes about an individual.

212. NSIs will need to consider how the inclusion of spatial data at every stage of the GSBPM is likely to impact on an organisation’s ability to prevent disclosure and the existing statistical disclosure control policies. There is a current shortage of research and documentation within the international statistical community on this topic and this is an area that will need further consideration to align with the rest of the activities that are taking place on the integration of statistics and geospatial information.

[36] http://ggim.un.org/docs/meetings/2nd%20UN-EG-ISGI/UN-GGIM%20EG%20Lisbon%20meeting%20session%204%20background%20paper%20terminology.pdf

[37] http://ggim.un.org/docs/meetings/2nd%20UN-EG-ISGI/UN-GGIM%20EG%20Lisbon%20meeting%20session%204%20background%20paper%20terminology.pdf

D. Disaggregation

213. The increased use of new data sources within the statistical process such as administrative and big data has the potential to cause problems for the production of statistics for small areas.

214. Census data is traditionally captured at the address level making it easy to aggregate up to small geographies for dissemination but data from these new data sources can often only be captured at higher geographic levels. This presents a problem for the modernisation of official statistics.

215. This is in conflict with the requirement of NSIs to publish their data at lower levels such as the SDG principle that “nobody gets left behind”. This means that despite data only being available at higher geographic levels there is a need to publish at smaller geographic regions to ensure this principle is adhered to.

216. For many of the proposed SDG indicators, and especially for analytical and national reporting purposes, geographic disaggregation will be required. For instance, the proposed indicator for target 4.2 is percentage of children under 5 years of age who are developmentally on track in health, learning and psychosocial well-being disaggregated by sex, location, wealth and others where data are available.

217. Methodologies already exist for providing disaggregated statistics (such as dasymetric disaggregation) but these methods are not used extensively and strongly rely on high quality and high resolution auxiliary information. The statistical community will need to consider what impact these methods will have on the quality of statistics compared to the current aggregation methods. NSIs will need to ensure that there is a coordinated approach between their teams providing statistics for monitoring sustainable development and those providing their geospatial services if the levels of disaggregation required for the SDGs is to be delivered.

E. Spatial Analysis

218. Statistical analysis is a well understood concept within the international statistical community. It is the process through which data can be interpreted and understood so that statistical reporting can support policy development.

219. Less well understood by the statistical community is spatial analysis as a tool for statistical reporting and NSIs face a challenge of how to interpret and understand data based on their location. Clustering of data based on location and bias towards urban or rural settlement patterns are all examples of how statistical data might be impacted by location.

220. There is at present, not only a lack of spatial analysis available within statistical reporting but a lack of geospatial information in general. Mapping almost exclusively consists of simple choropleths even where these are not the most appropriate mapping techniques and some statistics such as socio-economic data contain almost no geographic information below member state.

221. International initiatives such as UN-GGIM and GSBPM offer the opportunity to capture spatial analysis as a requirement but have so far failed to do so. The statistical community will therefore need to identify ways to grow the capability for spatial analysis and to formalise it within the frameworks of the statistical process.

VII. Conclusions and recommendations

222. There is an increasing recognition within both the statistical and geospatial communities of the requirement to integrate statistics with geospatial information to improve the quality and understanding of statistical data.

223. This increased recognition is reflected in the large number of international initiatives that are seeking to harmonise and understand how geospatial data can and does fit into the statistical process.

224. From the statistical perspective, the existing work of the Common Statistical Process Architecture (CSPA) and Generic Statistical Business Process Model (GSBPM) represent a useful framework for understanding how geospatial information can contribute to the modernisation of official statistics.

225. From the geographic perspective, initiatives such as United Nations Global Geospatial Information Management (UN-GGIM) are seeking to understand how they can change their own processes and data management to better make geospatial information available to statisticians.

226. These initiatives are giving statisticians access to new administrative and big data sources, and these new data sources provide their own challenges for statistics based around the levels at which they can be reported, disaggregation methods and disclosure control.

227. There is now an understanding that a single approach to the geographic dissemination of statistics isn't feasible given the differing requirements for statistical production and the advantages and disadvantages of the different geographic levels at which statistics can be published. Instead NSIs will need to consider how such a diverse set of geographies can all publish statistics without becoming disclosive.

228. The biggest challenge facing the statistical community is how to see all of the work being undertaken as a single cohesive approach to geospatial statistics. With so many stakeholders and initiatives focussing on this area the international statistical community needs to consider how all of this work can be consolidated and incorporated within the CSPA and GSBPM.

Geospatial information services based on official statistics

Working Paper Series on Statistics

The UNECE Working Paper Series on Statistics consists of studies prepared by leading experts in official statistics from the UNECE region. The Series presents and analyses timely topics in statistics and aims to identify emerging issues and share innovations. The studies often serve as a basis for launching new work to develop new statistics and guidelines.

This review is concerned with how geospatial information can be used in the production of official statistics. It will consider how geospatial information fits within the statistical process and what barriers exist to exploiting the value of geospatial information for official statistics.

Information Service
United Nations Economic Commission for Europe

Palais des Nations
CH - 1211 Geneva 10, Switzerland
Telephone: +41(0)22 917 44 44
Fax: +41(0)22 917 05 05
E-mail: info.ece@unece.org
Website: http://www.unece.org