Geospatial View of Generic Statistical Business Process Model

GeoGSBPM

(version 1.0, May 2021)

Acknowledgement

This paper was developed by the Geospatial Task Team under the Supporting Standards Group of the High-Level Group on the Modernisation of Official Statistics. The Task Team consisted of: Martin Brady (Australia), Tim Werschler (Canada), Kathrin Gebers (Germany), Essi Kaukonen and Antti Santaharju (Finland), Christophe Dzikowski (France), Zoltan Vereczkei (Hungary), Kevin McCormack (Ireland), Juan Munoz (Mexico), Anna Sławińska and Amelia Wardzińska-Sharif (Poland), Yulla Choi (Republic of Korea), Krishnan Ambady (United Kingdom), Jay Greenfield (Data Documentation Initiative), Cristian Fetie, Oliver Mueller, Hannes Reuter and Nikolaos Roubanis (Eurostat), Edgardo Greising (International Labour Organization), Mark Iliffe (United Nations Statistics Division), InKyung Choi, Taeke Gjaltema and Steven Vale (United Nations Economic Commission for Europe)

List of Abbreviations

API: Application Programming Interface

CES: Conference of European Statisticians

DGGS: Discrete Global Grid System

GAMSO: Generic Activity Model for Statistical Organisations

GeoDCAT: Geo Data Catalog Vocabulary

GeoGSBPM: Geospatial view of Generic Statistical Business Process Model

GeoJSON: Geo JavaScript Object Notation

GSGF: Global Statistical Geospatial Framework

GIS: Geographic Information System

GPS: Global Positioning System

GSBPM: Generic Statistical Business Process Model

HLG-MOS: High-Level Group for the Modernisation of Official Statistics

ICT: Information and Communications Technology

IGIF: Integrated Geospatial Information Framework

INSPIRE: Infrastructure for Spatial Information in the European Community

ISO: International Organization for Standardization

LOD: Linked Open Data

NGIA: National Geospatial Information Agency

NSDI: National Spatial Data Infrastructure

NUTS: Nomenclature of Territorial Units for Statistics

OGC: Open Geospatial Consortium

OWL: Web Ontology Language

PID: Persistent Identifier

RDF: Resource Description Framework

SDG: Sustainable Development Goal

SDMX: Statistical Data and Metadata eXchange

SKOS: Simple Knowledge Organization System

UN EG-ISGI: United Nations Expert Group on the Integration of Statistical and Geospatial Information

UNECE: United Nations Economic Commission for Europe

UN-GGIM: United Nations Experts on Global Geospatial Information Management

XML: eXtensible Mark-up Language

Table of Contents

1. Introduction **6**
Background 6
Situating the environment for the GSBPM and GSGF 7
Contextualising the geospatial view of the GSBPM 8
2. Geospatial-related activities and considerations **11**
Specify Needs Phase 11
Design Phase 14
Build Phase 18
Collect Phase 21
Process Phase 23
Analyse Phase 27
Disseminate Phase 29
Evaluate Phase 31
3. Overarching processes and corporate-level activities **33**
Strategic collaboration and cooperation 33
Metadata management 34
Quality management 35

1. Introduction

Background

1. The appropriate use of geospatial information is crucial in realising the full potential of the data produced in the data ecosystem. Geospatial information, as the digital currency of geographic location, is playing an increasingly important role for the work of statistical organisations. Primarily, authoritative geospatial information is produced by the National Geospatial Information Agencies (NGIAs) or mapping agencies. However, all data with a geographic location is a constituent component of the data ecosystem, which the national statistical organisations often finds itself as the custodian.

2. The data ecosystem in which statistical organisations operate is more diverse than ever, there are various actors, from government agencies, private companies to citizens, producing data with different tools and in different formats. With digitalisation and advance of technologies, data are also being generated by non-human agents at an explosive rate (e.g. sensor data, data from web-crawler, mobility data from cell phones). Although these datasets vary on "what" they are about and "how" they are generated, they can be linked through information on "where" the dataset is referring to. While unique unit identifiers such as personal ID and many social or demographic variables are considered confidential and difficult to derive from original datasets, such geospatial information is often available to a certain disaggregated level and this can be used to integrate data from different sources. Geospatial information is an unambiguous and universal key that cuts across all data. Statistical information combined with location information (henceforth referred to as "geospatially enabled statistics") can provide critical knowledge by the integration with other data produced by various actors in the data ecosystem to understand multi-faceted issues that the society currently faces such as sustainable development, rapid urbanisation and climate change.

3. Geospatially enabled statistics, in particular, at the sub-national and high spatial resolution, greatly increase the relevance of statistical information by providing the geographic context of the phenomenon that the dataset is capturing. This enables policy makers and researchers to more easily understand and analyse this geographic relationship, leading towards the development of more targeted, locally relevant, and actionable plans such as access to public infrastructure (e.g. school, transportation, green area), rural / urban inequality and emergency planning. The value of geospatially enabled statistics is not limited to the public sector. Wide use of map services through the web has lowered the access barrier to location information and changed the way it is used for decision making for all spheres of the society. Geospatially enabled statistics allow the end users of this data (including companies, enterprises, and citizens) to benefit from localised information more pertinent to their business and other needs. By allowing the geospatial analysis and the linkage with various data sets, geospatially enabled statistics can also open up the new research potential for the scientific and academic community.

4. Geography has long been understood as a fundamental component of the work of statistical organisations (e.g. in the geographical classification for designing sampling and processing raw data, as a base dimension with which statistical data are released, as a tool to support and plan field operations), yet the scope and extent of the usage has been limited. For example, the geographic granularity of statistical data is often released at a large regional level which makes it difficult to draw a meaningful geographic context but also not flexible enough to be integrated with other data sources. To address the information needs of various users in an increasingly complex and intertwined society, there is also a great need for statistical data to be geospatially enabled using consistent and common geographies, in an accessible and usable format.

5. It is important to note that geospatially enabled statistics are not just for particular use cases or one-off exercises. An understanding of the role and characteristics of geography is an important for all stages of the statistical production process and the production of geospatially enabled statistics should be a routine operation for statistical organisations. Further, as the novel coronavirus (COVID-19) global pandemic has highlighted, statistical organisations should be prepared to produce geospatially enabled statistical data in an efficient and timely manner. To ensure this occurs, geospatially relevant activities and considerations should be integrated into the regular production processes of statistical organisations, so that the design and production of geospatially enabled statistics can be conducted in a systematic and consistent way.

Situating the environment for the GSBPM and GSGF

6. To identify what activities and considerations are needed for the production of geospatially enabled statistics and document them in the context of the statistical production process, two global frameworks are used in this paper:

- The Generic Statistical Business Process Model (GSBPM)[1] describes the set of activities needed to produce official statistics. It provides a standard framework and harmonised terminology to help statistical organisations to modernise their statistical production processes. The GSBPM is one of the cornerstones of the standards-based modernisation strategy of the United Nations Economic Commission for Europe (UNECE) High-Level Group for the Modernisation of Official Statistics (HLG-MOS) and is widely adopted as a de-facto standard process model by the global official statistics community since its development in 2008. The model was endorsed by the Conference of European Statisticians (CES) in 2017;

- The Global Statistical Geospatial Framework (GSGF)[2] describes five Principles and supporting key elements for the production of harmonised and standardised geospatially enabled data. These five Principles are:
    - Principle 1: Use of fundamental geospatial infrastructure and geocoding;
    - Principle 2: Geocoded unit record data in a data management environment;
    - Principle 3: Common geographies for dissemination;
    - Principle 4: Statistical and geospatial interoperability;
    - Principle 5: Accessible and usable geospatially enabled statistics.

  Developed by the United Nations Expert Group on the Integration of Statistical and Geospatial Information (UN EG-ISGI), the GSGF was adopted by the United Nations Committee of Experts on Global Geospatial Information Management (UN-GGIM) in 2019 and endorsed by the Statistical Commission in 2020. GSGF is a key framework for facilitating the integration of statistical and geospatial information[3]. The resulting data,

[1] UNECE HLG-MOS "Generic Statistical Business Process Model" (https://statswiki.unece.org/display/GSBPM)
[2] UN-GGIM "Global Statistical Geospatial Framework" (http://ggim.un.org/meetings/GGIM-committee/9th-Session/documents/The_GSGF.pdf)
[3] Where the GSGF is the bridge between the statistical and geospatial communities, many NGIAs are guided by another key framework - the Integrated Geospatial Information Framework (IGIF). The IGIF provides a basis and guide for developing, integrating, strengthening and maximizing geospatial information management and related resources in all countries and offers countries an overarching framework that complements and strengthens their existing National Spatial Data Infrastructure (NSDI). Importantly, the IGIF is not just a

produced following the Principles, can be readily integrated with statistical, geospatial and other information to inform and facilitate data-driven and evidence-based decision making to support local, sub-national, national, regional, and global development priorities and agendas, such as the 2020 Round of Population and Housing Censuses and the 2030 Agenda for Sustainable Development. Several initiatives have been undertaken to support countries in adopting the Principles of the GSGF such as the Implementation Guide for the GSGF in Europe which was developed by the ESSnet project GEOSTAT 3[4] and is currently the subject of following work of GEOSTAT 4[5]. Further, the UN EG-ISGI continues its effort to provide practical and relevant guidance that enables countries to implement the GSGF in their national context.

7. As a standard process model in the statistical community, the GSBPM has an immediate connection to GSGF Principle 4 (Statistical and geospatial interoperability). Its common language and terminology can facilitate communication between the statistical and geospatial communities and provide a basis for understanding and aligning their business processes.

8. Further, the GSBPM can also be an enabling framework to help the GSGF Principles to be integrated into the production process of statistical organisations. The GSBPM lays out typical activities and steps that statistical organisations take when producing statistics and this provides a structure to document geospatial-related activities so that relevant actions are taken at the right stage of the production process. For example, consideration of common geographies for dissemination (GSGF Principles 3 and 5) should be taken into account from the early stage of the process. Ideally, the discussion about the type and the resolution of geographies and their implications should take place with users during the need specification stage, then reflected in the design stage and subsequently implemented in the process, analysis and dissemination stages according to these design decisions. This sequence of work can be modelled using GSBPM Phases and Sub-processes as building blocks

Contextualising the geospatial view of the GSBPM

9. The Geospatial View of the GSBPM (henceforth GeoGSBPM) describes geospatial-related activities, in particular, those that are needed to produce geospatially enabled statistics, using the framework of the GSBPM. Section 2 follows the structure of the eight GSBPM Phases and describes what activities and considerations should be included in each Phase. Section 3 discusses activities and considerations that should be done as overarching processes or at the corporate level to support the eight Phases of the production process[6]. These geospatial-related actions and considerations are identified while taking into account GSGF Principles so that the resulting statistics have a higher level of standardisation and geospatial flexibility, as well as a greater capacity for data integration. Although degree varies, each GSGF Principle is relevant to most GSBPM Phases and affects the production process through the overarching processes and corporate-level activities as depicted in Figure 1 below. Further, Table 1 provides a matrix of some of key activities that take place in the GSBPM Phase related to each GSGF Principle.

resource for NGIAs, but can help to strengthen the use of geospatial information within a country. For more about IGIF, see UN-GGIM IGIF (https://ggim.un.org/IGIF/).

[4] GEOSTAT 3 Project (https://www.efgs.info/geostat/geostat-3/)

[5] GEOSTAT 4 Project (https://www.efgs.info/geostat/geostat-4/)

[6] Note that corporate-level supporting activities are not in the scope of the GSBPM, but covered by the Generic Activity Model for Statistical Organisation (GAMSO), another HLG-MOS model complementing GSBPM. For more, see Section 3.

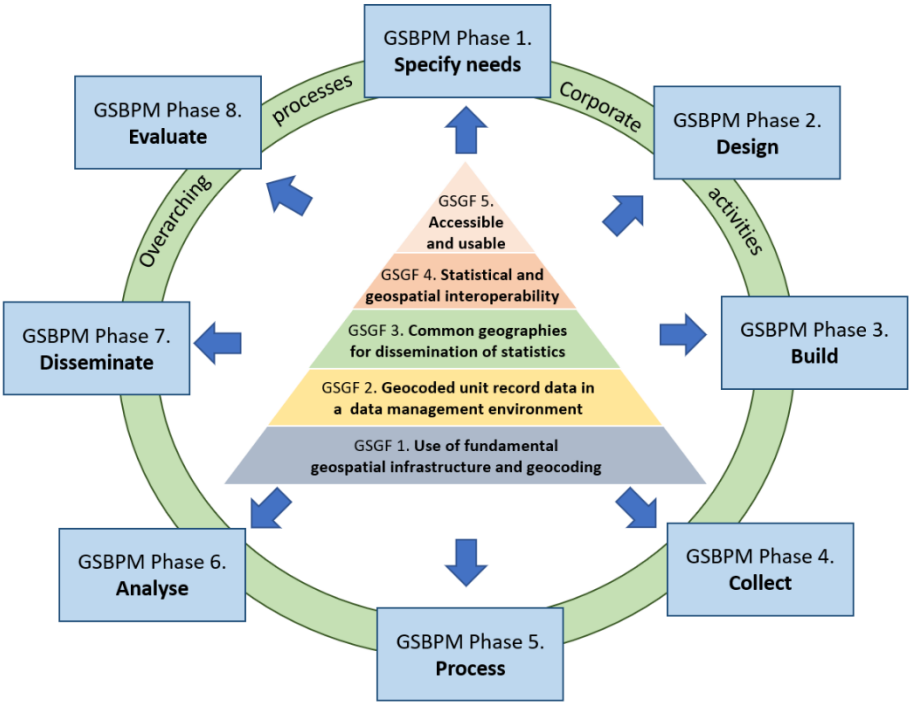

Figure 1. GSBPM and GSGF Principles

10. In addition to assisting the production of geospatially enabled statistics in a consistent and systematic way, the GeoGSBPM can support statistical organisations in the following ways:

- By identifying common activities required for the production of geospatially enabled statistics, it can facilitate sharing of geospatial services, methods and tools that can be applied regardless of data types, domains and output formats;
- By highlighting which geospatial-related activities and considerations are needed in the context of a typical statistical production process, it can assist efforts to make standards and technologies of the statistical and geospatial communities more interoperable;
- By clarifying the process in which statistical data and geospatial information flow and interact with each other, it can provide a common framework to manage quality and metadata of statistical and geospatial information and services.

Table 2. Geospatial-related activities in the GeoGSBPM and GSGF Principles

	GSBPM Phase Specify Needs	GSBPM Phase Design	GSBPM Phase Build	GSBPM Phase Collect	GSBPM Phase Process	GSBPM Phase Analyse	GSBPM Phase Disseminate	Overarching processes / Corporate activities
GSGF 1. Use of fundamental geospatial infrastructure and geocoding	When assessing data availability, the existence and availability of suitable geospatial information should be first identified from authoritative sources within the National Spatial Data Infrastructure (NSDI)	Geospatial variables (geographies) should be designed for the statistical unit level. Using point-based location as the base geospatial variable will provide considerable adaptability to changes over time and flexibility to aggregate up to various dissemination level geographies		Geocoding should be conducted for each statistical unit that is collected and at the most detailed level (e.g. point-based geocoding as opposed to area-based geocoding)	Standardisation should take place before the integration of datasets. It can be done through, for example, matching location information in the datasets with centralised standard systems (e.g. address matching, geocoding) which should be based on the national geospatial information context			Quality management include: identify the authoritative (external or internal) sources of reference data and establish quality profile of reference data
GSGF 2. Geocode unit record data in a data management environme nt		The design of components includes: point-of-entry validation for geographical information; matching strategy; and, spatial analysis			The mechanism of matching or geocoding the statistical unit-record established in Design phase should be consistently applied			Quality management include: develop quality dimensions and metrics to be used at different stages, and a consistent matching strategy
GSGF 3. Common geographies for production and dissemination of statistics	Needs of users in terms of geographies (e.g. size of unit, type) is discussed. Implications (e.g. cost, reliability, quality) should be communicated and consulted with users	When grid geographies are used, the choice of grid system should take existing regional and global systems into consideration		Inaccuracies in geospatial information detected during field collection should be documented and transferred to the central geospatial information system for maintenance and update if necessary (if permitted under statistical confidentiality rules)				
GSGF 4. Statistical and geospatial interoperability - data, standards and processes		Design of all production components should take into account standards used in the geospatial community	Geospatial services have a broad stakeholder group, statistical organisations should check and consult with service inventories of stakeholders before building components on their own			When preparing the analysis output, it is important to pay attention to semantic interoperability so that the output can be understood and used without ambiguities by users from different domains	International standards should be used as a norm to ensure that the products can be found and consumed easily across a range of various user groups from the public and private sectors	Alignment and harmonisation of geospatial metadata concepts with those of statistical metadata is critical
GSGF 5. Accessible and usable geospatially enabled statistics	Discussion on the output format is useful as users for high spatial resolution data (e.g. city, municipal authority) might require data to be provided in certain formats that are digestible within their GIS system. Implications of the size of geographic units in terms of confidentiality risk should be discussed with users	Design of these outputs should also take potential downstream uses into consideration. Accessibility and usability of geospatially enabled statistics and services can greatly increase by use of standards and open data formats	Metadata elements are put together during development of dissemination components so that they can be disseminated along with the data products and services. To make it more findable and accessible for both internal and external users, metadata should be documented using standard taxonomy and vocabulary			Cataloguing and tagging the content using relevant metadata standards can greatly increase the usability of the analysis outputs. Geospatial product components should be cross-checked with other components (e.g. tabular aggregates, before release so that they do not breach privacy on their own as well as in combination with other outputs		Statistical organisations are encouraged to explore the semantic web standards as a long-term strategic objective with successive milestones to achieve dissemination of data and metadata within the framework of Linked Open Data (LOD)

2. Geospatial-related activities and considerations

11. Section 2 follows the structure of the GSBPM and extends the original GSBPM text with descriptions of geospatial-related activities and considerations. While the focus is on the geospatially enabled statistics, the section also includes description of other related activities to emphasize the various roles of geospatial information in the statistical production in general. To distinguish the newly added texts, they are highlighted in grey in the remainder of Section 2. Note that the GeoGSBPM inherits the characteristics of the GSBPM such as non-linearity (i.e. steps do not need to be followed in a strict linear order) and applicability to various data sources[7].

Specify Needs Phase

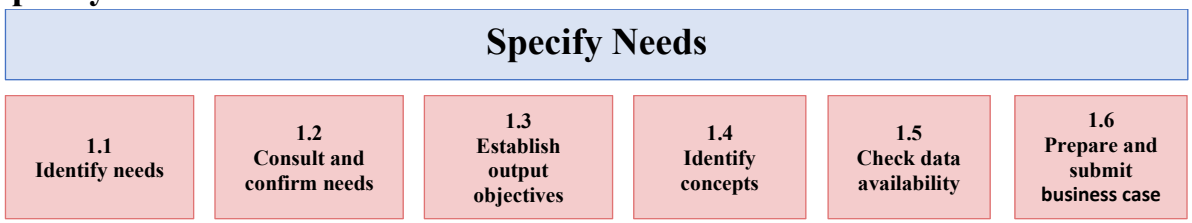

Figure 2. Specify needs phase and its sub-processes

12. This phase is triggered when a need for new statistics is identified or feedback about current statistics initiates a review. It includes all activities associated with engaging stakeholders to identify their detailed statistical needs (current or future), proposing high level solution options and preparing a business case to meet these needs.

13. Increasingly, users expect statistics to be geo-referenced by default and the referencing system to be more granular and flexible. As these geospatial needs can have a significant implication on the cost of production, reliability and the risk of privacy breaches, they have to be carefully examined from the very beginning of the production process.

14. The "Specify Needs" phase is broken down into six sub-processes (Figure 2), which are generally sequential, from left to right, but can also occur in parallel, and can be iterative. These sub-processes are:

1.1. Identify needs

15. This sub-process includes the initial investigation and identification of what statistics are needed and what is needed of the statistics. It may be triggered by a new information request or an environmental change such as a reduced budget. Action plans from evaluations of previous iterations of the process or from other processes might provide an input to this sub-process. It also includes consideration of practice amongst other (national and international) statistical organisations producing similar data and the methods used by those organisations.

1.2. Consult and confirm needs

16. This sub-process focuses on consulting with the internal and external stakeholders and confirming in detail the needs for the statistics. A good understanding of user needs is required so that the statistical organisation knows not only what it is expected to deliver, but also when, how, and, perhaps most importantly, why.

[7] For more, see GSBPM Section 2

17. Type of geography (e.g. administrative geographies, gridded geographies) and its advantages and disadvantages[8] (e.g. stability over time, comparability with other regions, ease of field verification) should be discussed with the users in this sub-process.

18. The smallest size of the geographic unit (e.g. NUTS level, grid cell) is also an important element to discuss in this sub-process. As the smaller the geography unit is, the more detailed the information becomes, users normally want the smallest geographic unit possible. However, high spatial resolution statistics often increase the production cost as well as the risk of privacy breaches. It could also affect the quality and reliability of the estimates produced as only few statistical units fall inside the small geographical area. The implication of the size of the geographic unit in terms of cost, quality and confidentiality should be communicated with the users.

19. For the second and subsequent iterations of this phase, the main focus will be on determining whether previously identified needs have changed. This detailed understanding of user needs is the critical part of this sub-process.

1.3 Establish output objectives

20. This sub-process identifies the statistical output objectives that are required to meet the user needs identified in sub-process 1.2 (Consult and confirm needs). It includes agreeing the suitability of the proposed outputs and their quality measures with users. It is also useful to discuss the format of the output in this sub-process, as users of high spatial resolution data (e.g. city, municipal authority) might require data to be provided in formats that can be readily integrated within their local enterprise GIS system. Legal frameworks (e.g. relating to confidentiality), and available resources are likely to be constraints when establishing output objectives.

1.4 Identify concepts

21. This sub-process clarifies the required concepts to be measured from the point of view of the users. At this stage, the concepts identified might not align with existing statistical standards. This alignment, and the choice or definition of the statistical and other concepts and variables to be used, takes place in sub-process 2.2 (Design variable descriptions).

1.5 Check data availability

22. This sub-process checks whether current sources of data could meet user requirements and the conditions under which they would be available including any restrictions on their use. An assessment of possible alternatives would normally include research into potential administrative or other non-statistical sources of data, to:

    - Determine whether they would be suitable for use for statistical purposes (e.g. the extent to which administrative concepts match data requirements, timeliness and quality of the data, security and continuity of data supply);
    - Assess the division of responsibilities between data providers and the statistical organisation;

[8] Administrative geographies are commonly used as they correspond to boundaries that users are familiar with but can be considered unstable as they change over time. Gridded geographies are increasing popular for their flexibility. For more about administrative and gridded geographies, see UN-GGIM "GSGF Annex B: Standards, Quality and Enabling Frameworks".

- Check necessary ICT resources (e.g. data storage, technology required to handle incoming data and data processing) as well as any formal agreements with data providers for accessing and sharing the data (e.g. formats, delivery, accompanying metadata and quality check).

23. During the data assessment, the availability of the geospatial information associated with the data should be checked at the data source. Geospatial information may exist at the point-based level (e.g. x-y coordinates) or at the coarse area-based level (e.g. administrative boundary) and this should be compared with geospatial requirements that users specified (e.g. geography type, size, date of reference of the data, availability of time series). If geospatial information does not exist or meet user requirements, it could be obtained separately from other sources (e.g. building register) or derived using auxiliary data (e.g. electricity usage data or satellite data can be used to derive location of dwellings). It is crucial that geospatial information should be first checked and obtained from authoritative sources of the National Spatial Data Infrastructure (NSDI) to the extent possible. Availability of metadata (e.g. time stamp, geographical classification) for the geospatial information should be also checked as it is important to ensure comparability among different data sets.

24. When existing sources have been assessed, a strategy for filling any remaining gaps in the data requirement is prepared. This may include identifying possible partnerships with data holders. This sub-process also includes a more general assessment of the legal framework in which data would be collected and used, and may therefore identify proposals for changes to existing legislation or the introduction of a new legal framework.

1.6 Prepare and submit business case

25. This sub-process documents the findings of the other sub-processes in this phase in the form of a business case to get approval to implement the new or modified statistical business process. Such a business case would need to conform to the requirements of the approval body, but would typically include elements such as:

    - A description of the "As-Is" business process (if it already exists), with information on how the current statistics are produced, highlighting any inefficiencies and issues to be addressed;
    - The proposed "To-Be" solution, detailing how the statistical business process will be developed to produce the new or revised statistics;
    - An assessment of costs and benefits, as well as any external constraints.

26. The business case describes options and makes recommendations. It may include the benefits, costs, deliverables, time frame, budget, required technical and human resources, risk assessment and impact on stakeholders for each option. For high-resolution statistics with small geography unit, analysis of privacy concerns and plans how to address them are included in the risk assessment.

27. After the business case is prepared, it is submitted for approval to move to the next phase of the business process. At this sub-process, a “go”/“no go” decision is made. Typically, the business case is reviewed and formally approved or disapproved by the appropriate sponsors and governance committees.

Design Phase

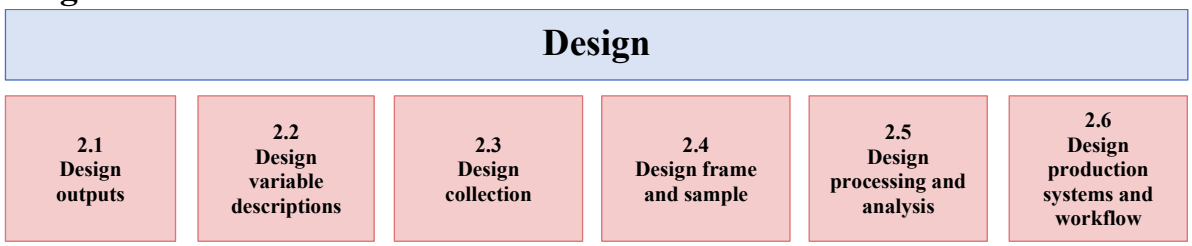

Figure 3. Design phase and its sub-processes

28. This phase describes the development and design activities, and any associated practical research work needed to define the statistical outputs, concepts, methodologies, collection instruments and operational processes. It includes all the design elements needed to define or refine the statistical products or services identified in the business case. This phase specifies all relevant metadata, ready for use later in the business process, as well as quality assurance procedures. For statistical outputs produced on a regular basis, this phase usually occurs for the first iteration and whenever improvement actions are identified in the "Evaluate" phase of a previous iteration.

29. Design activities make substantial use of international and national standards in order to reduce the length and cost of the design process, and enhance the comparability and usability of outputs. Organisations are encouraged to reuse or adapt design elements from existing processes, and to consider geospatial aspects of data in the design to enhance the usability and value of the statistical information. Additionally, outputs of design processes may form the basis for future standards at the organisational, national or international levels.

30. Standards used for data and services in the geospatial community may be different from those of statistical community[9]. It is therefore important for geospatially enabled statistics that each phase of production (e.g. collection, processing, disseminating) is designed in a way to ensure interoperability between the two fields.

31. The "Design" phase is broken down into six sub-processes (Figure 3), which are generally sequential, from left to right, but can also occur in parallel, and can be iterative. These sub-processes are:

2.1 Design outputs

32. This sub-process contains the detailed design of the statistical outputs, products and services to be produced, including the related development work and preparation of the systems and tools used in the "Disseminate" phase. Processes governing access to any confidential outputs are also designed here. When small geographic units are used for dissemination, there is a higher chance of statistical units being unintentionally identified. Geospatially enabled statistics also have a

[9] For example, geospatial data often use a vector or raster data model which requires data format and metadata different from those used in the statistical community.

particular risk of confidentiality breach, for example, by geographic differencing[10]. These risks should be considered in the design to ensure that the output can be released with confidence[11].

33. Outputs should be designed to follow existing standards wherever possible, so inputs to this process may include metadata from similar or previous collections (including extractions from statistical, administrative, geospatial and other non-statistical registers and databases), international standards, and information about practices in other statistical organisations from sub-process 1.1 (Identify needs). Outputs may also be designed in partnership with other interested bodies, particularly if they are considered to be joint outputs, or they will be disseminated by another organisation.

34. Spatial visualisation can present data in a more intuitive way and can reveal patterns that are not easily detectable otherwise (e.g. through tabular data). Geographic viewing tools are a powerful way to help users to understand the geographic context of the issues that they are trying to solve with the data. Therefore, it is recommended to include spatial visualisation and GIS services components in the output in addition to traditional formats (e.g. table, chart). Design of these outputs should also take potential downstream usages into the consideration when deciding output formats (e.g. providing maps in an editable format rather than a static image file). Accessibility and usability of geospatially enabled statistics and services can greatly increase by the use of standards and open data formats (e.g. XML, GeoJSON).

2.2 Design variable descriptions

35. This sub-process defines the variables to be collected via the collection instrument, as well as any other variables that will be derived from them in sub-process 5.5 (Derive new variables and units), and any statistical or geospatial-classifications that will be used. It is expected that existing national and international standards will be followed wherever possible.

36. Geospatial variables (geographies) that are used while collecting data at a statistical unit level are not usually the same as those that are used for dissemination. Hence, they should be designed at the statistical unit level using point-based location[12] as the base geospatial variable, as it will provide a considerable adaptability to changes over time and flexibility to aggregate up to various dissemination-level geographies. For gridded geographies, it is important to use a grid system that is comparable with the existing regional or global grid system (e.g. Discrete Global Grid System (DGGS)[13]) as it will greatly increase usability of the output. Different types of grid (e.g. hexagon,

[10] Geographic differencing is the process where the same statistical data is obtained for two similarly shaped regions and the data from one region is subtracted from the other larger region. By using this method, it is possible to obtain data for the area that is not common to both regions which might result in a privacy breach (source: Australian Bureau of Statistics "Statistical Spatial Framework Guidance Material – Protecting Privacy for Geospatially Enabled Statistics: Geographic Differencing" (https://www.abs.gov.au/websitedbs/d3310114.nsf/home/Statistical+Spatial+Framework+Guidance+Material/$File/Protecting+Privacy.pdf)).

[11] For more about privacy issues related to geospatial statistics and recommendations to ensure confidentiality, see UN-GGIM "GSGF: Implementing Privacy and Confidentiality", a background document to the eleventh session of UN-GGIM (to be made available).

[12] For more about point-based foundation for statistics, see GEOSTAT2 "A Point-based Foundation for Statistics" (https://www.efgs.info/geostat/geostat2/).

[13] Developed under Open Geospatial Consortium (OGC), DGGS represents the Earth with a tessellation of nested cells which allows rapid assembly and enables data integration without the difficulties of different coordinate reference system (source: OGC DGGS Standards Working Group (https://www.ogc.org/projects/groups/dggsswg)).

rectangular) and their advantages and disadvantages can be assessed when designing gridded geographies.

37. This sub-process may need to run in parallel with sub-process 2.3 (Design collection), as the definition of the variables to be collected, and the choice of collection instruments may be inter-dependent to some degree. Preparation of metadata descriptions of collected and derived variables, statistical and geospatial classification is a necessary precondition for subsequent phases.

2.3 Design collection

38. This sub-process determines the most appropriate collection instruments and methods which may depend on the type of data collection (census, sample survey, or other), the collection unit type (enterprise, person, or other) and the available sources of data. The actual activities in this sub-process will vary according to the type of collection instrument required, which can include computer assisted interviewing, paper questionnaires, administrative registers (e.g. by using existing service interfaces), data transfer methods, web-scraping technologies as well as technology for geospatial information. Direct or indirect use of administrative data may be introduced in the data collection mode for either controlling survey data or assisting it when capturing survey information.

39. This sub-process includes the design of the collection instruments, questions and response templates (in conjunction with the variables and statistical classifications designed in sub-process 2.2 (Design variable descriptions)). It also includes the confirmation of any formal agreements. This sub-process is enabled by tools such as question libraries (to facilitate the reuse of questions and related attributes), questionnaire tools (to enable the quick and easy compilation of questions into formats suitable for cognitive testing) and agreement templates (to help standardise terms and conditions). This sub-process also includes the design of provider management systems that are specific to this business process.

40. Where statistical organisations do not collect data directly (i.e. a third party controls the collection and processing of the data), this sub-process may include the design of mechanisms to monitor the data and the metadata to assess impacts of any change made by the third party.

41. Geospatial information can be obtained in different ways depending on the data collection mode. When field collection is involved, geospatial information of the statistical unit (e.g. address, coordinate) may already be provided if the sampling frame is geo-referenced. If such information is not available or deemed inaccurate, the geospatial information of the statistical unit may need to be captured during the field operation, for example, by means of a sensor (GPS) in the collection device or manually on a digital map by surveyors in the case of remote regions. For surveys where the main phenomenon of interest is of a geospatial nature (e.g. passenger mobility survey, road freight survey) or when data are collected through sensors, GPS coordinates can be registered continuously and automatically via collection devices. It is important to make sure that the collection device is designed to record necessary metadata (e.g. GIS system version, geodesic reference system).

42. When geospatial information is collected along with data (during surveys or data transfer from third parties), the efficient and sustainable way to ensure quality is to make sure the information is

accurate from the source[14]. Therefore, design of the collection instrument should include a point-of-entry validation tool.

2.4 Design frame and sample

43. This sub-process only applies to processes which involve data collection based on sampling, such as through statistical surveys. It identifies and specifies the population of interest, defines a sampling frame (and, where necessary, the register from which it is derived), and determines the most appropriate sampling criteria and methodology (which could include complete enumeration). Common sources for a sampling frame are administrative and statistical registers, censuses and information from other sample surveys. It may include geospatial information and classifications.

44. This sub-process describes how these sources can be combined if needed. Analysis of whether the frame covers the target population should be performed. A sampling plan should be made. The actual sample is created in sub-process 4.1 (Create frame and select sample), using the methodology specified in this sub-process.

45. Geospatial information plays an important role in designing frames and samples. Geo-referenced frames can help reduce the survey cost (e.g. optimising routes for data collection) and ensure the geographical representativeness in the sample. Geo-sampling (i.e. sampling that takes geospatial distribution into consideration) can increase the efficiency of estimates in spatial analysis.

2.5 Design processing and analysis

46. This sub-process designs the statistical processing methodology to be applied during the "Process" and "Analyse" phases. This can include among others, specification of routines and rules for coding, editing and imputation which may vary based on the mode of data collection and source of data.

47. Similar to statistical data, geospatial information obtained during the collection phase will undergo a range of processing (e.g. editing, imputation, validation)., in particular, those from the third party as different sources might store the geospatial information in different ways. Geospatial information is often used as a key variable for integrating data from various sources, hence, the standardisation of geospatial information to enable its use across the production process and the understanding of its quality are critical. Matching and non-matching strategy for integration and record linkage is developed in this sub-process. This sub-process includes design of processing methodologies specifically needed for geospatial information (e.g. point-in-polygon processing) as well as other geospatial services.

48. This sub-process also includes design of specifications for data integration from multiple data sources, validation of data and estimation Statistical disclosure control methods are also designed here if they are specific to this business process.

49. Geospatially enabled statistics, in particular at a high geospatial spatial resolution, can allow statistical organisations to produce analytical outputs at more disaggregated levels and conduct a

[14] For more about the use of point-of-entry validation in collection, see Requirement 2.5 of GEOSTAT 3 "GSGF Europe – Implementation Guide for GSGF in Europe" (https://www.efgs.info/wp - content/uploads/geostat/3/GEOSTAT3_GSGF_EuropeanImplementationGuide_v1.0.pdf).

wide range of spatial analysis[15] (e.g. map visualisation, spatial-temporal regression). Design of such analysis and analytical output can be conducted in this sub-process.

2.6 Design production systems and workflow

50. This sub-process determines the workflow from data collection to dissemination, taking an overview of all the processes required within the whole production process and ensuring that they fit together efficiently with no gaps or redundancies. Various systems and databases are needed throughout the process. The GSBPM can be used as the basis of the business architecture layer when a statistical organisation has an existing enterprise architecture in place. The design might be adjusted to fit the organisation. A general principle is to reuse processes and technology across many statistical business processes, so existing production solutions (e.g. services, systems and databases) should be examined first, to determine whether they are fit for purpose for this specific production process, then, if any gaps are identified, new solutions should be designed. This sub-process also considers how staff will interact with systems and who will be responsible for what and when.

Build Phase

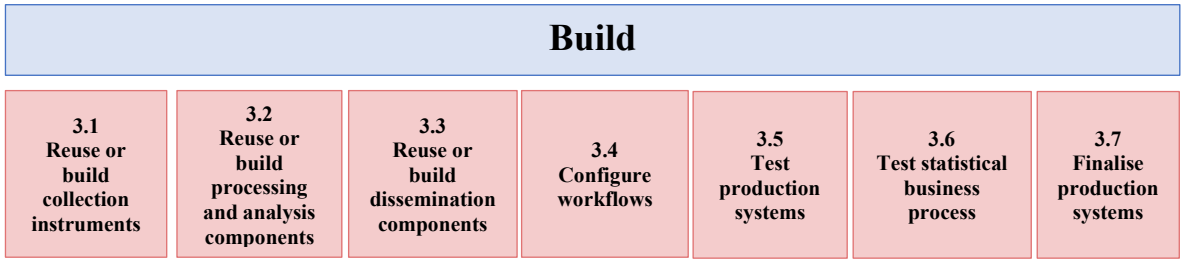

Figure 4. Build phase and its sub-processes

51. This phase builds and tests the production solution to the point where it is ready for use in the "live" environment. The outputs of the "Design" phase are assembled and configured in this phase to create the complete operational environment to run the process. New services are built by exception, created in response to gaps in the existing catalogue of services sourced from within the organisation and externally. These new services are constructed to be broadly reusable in alignment with the business architecture of the organisation where possible.

52. While statistical services are often used mainly within statistical organisations, geospatial information services often have a much broader stakeholder group. For example, services for the search of geospatial information or correcting misspelled addresses can be used not only by statistical organisations, but also by administrative data providers, NGIAs, and others. Therefore, geospatial information and their related services may already exist in the inventory of these stakeholders and should be checked to avoid duplication of efforts before building the components within the statistical organisations. If and when a new geospatial service is needed, it should be built in collaboration with NGIA and other stakeholders.

53. For statistical outputs produced on a regular basis, this phase usually occurs for the first iteration, following a review or a change in methodology or technology, rather than for every iteration.

[15] For more about the spatial analysis for official statistics, see National Institute of Statistics and Economic Studies (Insee) of France “Handbook of Spatial Analysis” (https://www.insee.fr/en/information/3635545).

54. The “Build” phase is broken down into seven sub-processes (Figure 4), which are generally sequential, from left to right, but can also occur in parallel, and can be iterative. The first three sub-processes are concerned with the development and improvement of systems used in collection, processing, analysis and dissemination of data. The last four sub-processes focus on the end-to-end process. These sub-processes are:

3.1 Reuse or build collection instruments

55. This sub-process describes the activities to build and reuse the collection instruments to be used during the "Collect" phase. The collection instruments are built based on the design specifications created during the "Design" phase. A collection may use one or more modes to receive the data (e.g. personal or telephone interviews; paper, electronic or web questionnaires; SDMX web services; automatic location collection tools). Collection instruments may also be data extraction routines used to gather data from existing statistical or administrative registers (e.g. by using existing service interfaces).

56. This sub-process also includes preparing and testing the contents and functioning of that collection instrument (e.g. cognitive testing of the questions in a questionnaire). It is recommended to consider the direct connection of collection instruments to a metadata system, so that metadata can be more easily captured in the collection phase. Connecting metadata and data at the point of capture can save work in later phases. Capturing the metrics of data collection (paradata) is also an important consideration in this sub-process for calculating and analysing process quality indicators.

3.2 Reuse or build processing and analysis components

57. This sub-process describes the activities to reuse existing components or build new components needed for the “Process” and “Analyse” phases, as designed in the "Design" phase. Services may include dashboard functions and features, information services, transformation functions, geospatial information services, workflow frameworks, provider and metadata management services.

3.3 Reuse or build dissemination components

58. This sub-process describes the activities to build new components or reuse existing components needed for the dissemination of statistical products as designed in sub-process 2.1 (Design outputs). All types of dissemination components are included, from those that produce traditional paper publications to those that provide web services, (linked) open data outputs, geospatial statistics, maps (static and interactive), or access to microdata.

59. Dissemination of geospatially enabled statistics requires additional metadata (e.g. geodesic system) for users to accurately understand and use the data. Metadata elements that are disseminated with data products and services need to be put together during development of dissemination components. To make it more findable and accessible for both internal and external users, metadata should be documented using a standard taxonomy and vocabulary (e.g. ISO 19119, GeoDCAT).

60. Safeguards and protections against the risk of disclosing individual information, in particular, for high resolution geospatial statistics, should be built and tested in this sub-process.

3.4 Configure workflows

61. This sub-process configures the workflow, systems and transformations used within the business processes, from data collection through to dissemination. In this sub-process, the workflow is configured based on the design created in sub-process 2.6 (Design production systems and workflows). This could include modifying a standardised workflow for a specific purpose, assembling the workflows for the different phases together (possibly with a workflow/business process management system) and configuring systems accordingly.

3.5 Test production systems

62. This sub-process is concerned with the testing of assembled and configured services and related workflows. It includes technical testing and sign-off of new programmes and routines, as well as confirmation that existing routines from other statistical business processes are suitable for use in this case. Whilst part of this activity concerning the testing of individual components and services could logically be linked with sub-process 3.1, 3.2 and 3.3, this sub-process also includes testing of interactions between assembled and configured services, and ensuring that the whole production solution works in a coherent way.

3.6 Test statistical business process

63. This sub-process describes the activities to manage a field test or pilot of the statistical business process. Typically, it includes a small-scale data collection, to test the collection instruments, followed by processing and analysis of the collected data, to ensure the statistical business process performs as expected. Following the pilot, it may be necessary to go back to a previous step and make adjustments to collection instruments, systems or components. For a major statistical business process, e.g. a population census, there may be several iterations until the process is working satisfactorily.

3.7 Finalise production systems

64. This sub-process includes the activities to put the assembled and configured processes and services, including modified and newly-created services, into production ready for use. The activities include:

    - Producing documentation about the process components, including technical documentation and user manuals;
    - Training the users on how to operate the process;

65. Moving the process components into the production environment and ensuring they work as expected in that environment (this activity may also be part of sub-process 3.5 (Test production system)).

Collect Phase

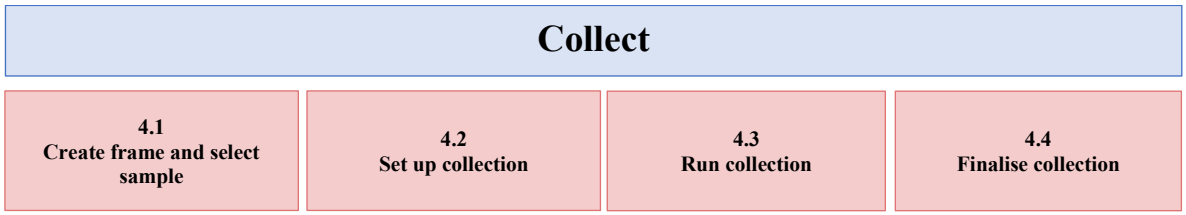

Figure 5. Collect phase and its sub-processes

66. This phase collects or gathers all necessary information (e.g. data, metadata and paradata), using different collection modes (e.g. acquisition, collection, extraction, transfer), and loads them into the appropriate environment for further processing. Whilst it can include validation of data set formats, it does not include any transformations of the data themselves, as these are all done in the "Process" phase. For statistical outputs produced regularly, this phase occurs in each iteration.

67. The "Collect" phase is broken down into four sub-processes (Figure 5), which are generally sequential, from left to right, but can also occur in parallel, and can be iterative. These sub-processes are:

4.1 Create frame and select sample

68. This sub-process establishes the frame and selects the sample for this iteration of the collection, as specified in sub-process 2.4 (Design frame and sample). It also includes the coordination of samples between instances of the same business process (e.g. to manage overlap or rotation), and between different processes using a common frame or register (e.g. to manage overlap or to spread response burden). Quality assurance and approval of the frame and the selected sample are also undertaken in this sub-process, though maintenance of underlying registers, from which frames for several statistical business processes are drawn, is treated as a separate business process. The sampling aspect of this sub-process is not usually relevant for processes based entirely on the use of pre-existing sources (e.g. administrative registers, web sites) as such processes generally create frames from the available data and then follow a census approach. Variables from administrative and other non-statistical sources of data can be used as auxiliary variables in the construction of sampling design.

4.2. Set up collection

69. This sub-process ensures that the people, processes and technology (e.g. web-based applications, GPS system) are ready to collect data and metadata, in all modes as designed. It takes place over a period of time, as it includes the strategy, planning and training activities in preparation for the specific instance of the statistical business process. Where the process is repeated regularly, some (or all) of these activities may not be explicitly required for each iteration. For one-off and new processes, these activities can be lengthy. For survey data, this sub-process includes:

- Preparing a collection strategy (e.g. optimising routes for data collection using geospatial information);
- Training collection staff;
- Training system using supervised machine learning techniques;
- Ensuring collection resources are available (e.g. laptops, collection apps, APIs);
- Agreeing on terms with any intermediate collection bodies (e.g. sub-contractors for computer assisted telephone interviewing, web services);

- Configuring collection systems to request and receive the data;
- Ensuring the security of data to be collected;
- Preparing collection instruments (e.g. printing questionnaires, pre-filling them with existing data, loading questionnaires and data onto interviewers' computers, APIs, web scraping tools);
- Ensuring GIS tools are configured appropriately for the specific collection instance;
- Providing information for respondents (e.g. drafting letters or brochures explaining the purpose of the survey, notifying respondents when online reporting instruments will be made available);
- Translating of materials (e.g. into the different languages spoken or used in the country).

70. For non-survey sources, this sub-process ensures that the necessary processes, systems and confidentiality procedures are in place, to receive or extract the necessary information from the source. This includes:

- Evaluating requests to acquire the data and logging the request in a centralised inventory;
- Initiating contacts with organisations providing the data, and sending an introductory package with details on the process of acquiring the data;
- Checking detailed information about files and metadata with the data provider and receiving a test file to assess if data are fit for use;
- Arranging secure channels for the transmission of the data.

4.3. Run collection

71. This sub-process is where the collection is implemented. The different collection instruments are used to collect or gather the information which may include raw microdata or aggregates produced at the source, as well as any associated metadata. It can include the initial contact with providers and any subsequent follow-up or reminder actions. It may include manual data entry at the point of contact, or fieldwork management, depending on the source and collection mode. It records when and how providers were contacted, and whether they have responded. Depending on the geographical frame and the technology used, geo-coding may need to be done at the same time as collection of the data by using inputs from GPS systems, putting a mark on a map, etc. When geocoding[16] is carried out during collection, it should be done for each statistical unit that is collected and at the most detailed level (e.g. point-based geocoding as opposed to area-based geocoding). This sub-process also includes the management of the providers involved in the current collection, ensuring that the relationship between the statistical organisation and data providers remains positive, and recording and responding to comments, queries and complaints. Proper communication with reporting units and minimisation of the number of non-respondents contribute significantly to a higher quality of the collected data.

72. For administrative, geographical or other non-statistical data, the provider is either contacted to send the information or sends it as scheduled. This process may be time consuming and might require follow-ups to ensure that data and metadata are provided according to the agreements. In

[16] Geocoding can be considered as a method of linking a description of a location to the location's measurable position in space. Geocoding links unreferenced location information (e.g., an address, or other location description) associated with a statistical unit (e.g., housing unit or business) to a set of coordinates within a coordinate system (source: UN-GGIM "GSGF: Implementing Geocoding" (http://mdgs.un.org/unsd/statcom/52nd-session/documents/BG-4j-EG-ISGI_Scoping_Paper-on_Geocoding-E.pdf)) or any other geocode system established within the organisations (e.g. mesh block code).

the case where the data are published under an Open Data license and exist in machine-readable form, they may be freely accessed and used.

73. This sub-process may also include the monitoring of data collection (e.g. real time monitoring of location of interviewers using geospatial information) and making any necessary changes to improve data quality. This includes generating reports, visualising and adjusting the acquisition process to ensure the data are fit for use. When the collection meets its targets, it is closed and a report on the collection is produced. Some basic checks of the structure and integrity of the information received may take place within this sub-process, (e.g. checking that files are in the right format and contain the expected fields).

4.4. Finalise collection

74. This sub-process includes loading the collected data and metadata into a suitable electronic environment for further processing. It may include manual or automatic data capture, for example, using clerical staff or optical character recognition tools to extract information from paper questionnaires, or converting the formats of files or encoding the variables received from other organisations. It may also include analysis of the metadata and paradata associated with collection to ensure the collection activities have met requirements. In cases where there is a physical collection instrument, such as a paper questionnaire, which is not needed for further processing, this sub-process manages the archiving of that material. When the collection instrument uses software such as an API or an app, this sub-process also includes the versioning and archiving of these. When inaccuracies of geospatial information are detected during the field collection (e.g. new settlement or district), this information should be documented and used to update the geospatial information systems of the statistical organisation, as it can affect downstream tasks (e.g. geospatial classification, estimates based on territory). Subject to maintaining statistical confidentiality, some corrections may also be fed back to mapping and cadastral agencies where appropriate.

Process Phase

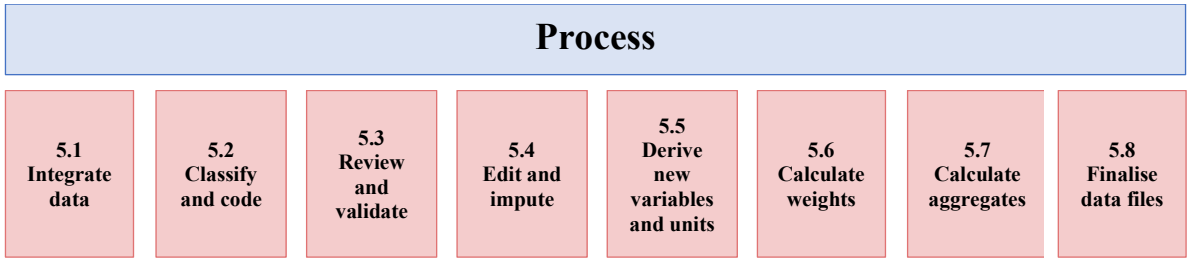

Figure 6. Process phase and its sub-processes

75. This phase describes the processing of input data and their preparation for analysis. It is made up of sub-processes that integrate, classify, check, clean, and transform input data, so that they can be analysed and disseminated as statistical outputs. While specific methodologies may differ, geospatial variables go through similar steps of processing as statistical variables. Depending on the role of the geospatial variable, however, it may be recommended to process geospatial variables before other variables (e.g. when location information is used as basis for data integration). For statistical outputs produced regularly, this phase occurs in each iteration. The sub-processes in this phase can apply to data from both statistical and non-statistical sources (with the possible exception of sub-process 5.6 (Calculate weights), which is usually specific to survey data).

76. The "Process" and "Analyse" phases can be iterative and parallel. Analysis can reveal a broader understanding of the data, which might make it apparent that additional processing is needed. Sometimes the estimates being processed might be already published aggregates (undertaken according to a Revision Policy).

77. Activities within the "Process" and "Analyse" phases may also commence before the "Collect" phase is completed. This enables the compilation of provisional results where timeliness is an important concern for users, and increases the time available for analysis.

78. The “Process” phase is broken down into eight sub-processes (Figure 6), which may be sequential, from left to right, but can also occur in parallel, and can be iterative. These sub-processes are:

5.1. Integrate data

79. This sub-process integrates data from one or more sources. It is where the results of sub-processes in the "Collect" phase are combined. The input data can be from a mixture of external or internal sources, and a variety of the collection instruments, including extracts of administrative and other non-statistical data sources. Administrative data or other non-statistical sources of data can substitute for all or some of the variables directly collected from survey. This sub-process also includes harmonising or creating new figures that agree between sources of data. The result is a set of linked data. Data integration can include:

    - Combining data from multiple sources, as part of the creation of integrated statistics such as national accounts;
    - Combining geospatial data and statistical data or other non-statistical data (e.g. combining satellite data on crop fields with agricultural statistics, combining road network data with population statistics);
    - Data pooling, with the aim of increasing the effective number of observations of some phenomena;
    - Matching or record linkage routines, with the aim of linking micro or macro data from different sources;
    - Data fusion - integration followed by reduction or replacement;
    - Prioritising, when two or more sources contain data for the same variable, with potentially different values.

80. When combining data from different sources, the geographic units used in the data might be different. The matching strategy for the geographic unit should be consistently applied (e.g. how a grid unit of population data is determined to be inside an administrative boundary in agricultural data) and any non-matching should be document with quality measures as developed in Phase 2 Design.

81. Geospatial information (e.g. address, x- and y-coordinate, or a geographical name) can play an important role in bringing together information from various domains by enabling integration of datasets from different sources using the location information as a matching key variable (e.g. integrating administrative data with survey data using address or postal code that exists in both datasets). To ensure the quality of the integration, standardising the geospatial information in the different datasets is critical. This standardisation would normally take place before the integration of datasets and can be done through, for example, matching location information in the datasets

with a centralised standard system (e.g. address matching, geocoding)[17] which should be part of the national spatial data infrastructure. This linkage, ideally done through consistent, unambiguous and persistent identifiers (PIDs), can also allow the dataset to use various additional geospatial information within the address registry or geocode database. In the absence of such a system, organisations may rely on other ways to reference location (e.g. GPS coordinates), alternative sources (e.g. address registry from utility provider) or higher-level geography (e.g. large geographical area).

82. Data integration may take place at any point in this phase, before or after any of the other sub-processes. There may also be several instances of data integration in any statistical business process. Following integration, depending on data protection requirements, data may be de-identified, that is stripped of identifiers such as name and address, to help to protect confidentiality.

5.2. Classify and code

83. This sub-process classifies and codes the input data. For example, automatic (or clerical) coding routines may assign numeric codes to text responses according to a pre-determined statistical classification to facilitate data capture and processing. Some questions have coded response categories on the questionnaires or administrative source of data, others are coded after collection using an automated process (which may apply machine learning techniques) or an interactive, manual process. When geocoding is conducted, it should be done at the unit record level before unit derivation (sub-process 5.5) or aggregation (sub-process 5.7) using the smallest geography available.

5.3. Review and validate

84. This sub-process examines data to identify potential problems, errors and discrepancies such as outliers, item non-response and miscoding. It can also be referred to as input data validation. It may be run iteratively, validating data against pre-defined edit rules, usually in a set order. It may flag data for automatic or manual inspection or editing. Reviewing and validating can apply to data from any type of source, before and after integration, as well as imputed data from sub-process 5.4 (Edit and impute). Whilst validation is treated as part of the "Process" phase, in practice, some elements of validation may occur alongside collection activities, particularly for modes such as computer assisted collection. Whilst this sub-process is concerned with detection and localisation of actual or potential errors, any correction activities that actually change the data is done in sub-process 5.4 (Edit and impute)

5.4. Edit and impute

85. Where data are considered incorrect, missing, unreliable or outdated, new values may be inserted or outdated data may be removed in this sub-process. The terms editing and imputation cover a variety of methods to do this, often using a rule-based approach. Specific steps typically include:

[17] The address matching and geocoding can be seen as data integration on their own (e.g. matching address in a dataset with valid address in an address register or geocode database). However, in this paper, the term "data integration" is used for the datasets that are of primary interest for the current business process and the address matching or geocoding are considered as a part of the standardisation or coding activity for this integration to take place. Note also that address matching and geocoding can be modelled as a process that uses a chain of sub-processes in Phase 5 (e.g. matching and linking address in survey dataset with geocoded address (sub-process 5.1), assigning the corresponding geocode (sub-process 5.2) while iteratively reviewing the address (sub-process 5.3) and editing / imputing (sub-process 5.4) where necessary.

- Determining whether to add or change data;
- Selecting the method to be used;
- Adding/changing data values;
- Writing the new data values back to the data set, and flagging them as changed;
- Producing metadata on the editing and imputation process.

5.5. Derive new variables and units

86. This sub-process derives data for variables and units that are not explicitly provided in the collection, but are needed to deliver the required outputs. It derives new variables by applying arithmetic formulae to one or more of the variables that are already present in the dataset, or applying different model assumptions. This activity may need to be iterative, as some derived variables may themselves be based on other derived variables. It is therefore important to ensure that variables are derived in the correct order. New units may be derived by aggregating or splitting data for collection units, or by various other estimation methods. Examples include deriving households where the collection units are persons or enterprises where the collection units are legal units. Point-based location information provided with unit record data can make the derivation of new geographical variables and conversion to different geographical variables flexible and straightforward.

5.6. Calculate weights

87. This sub-process creates weights for unit data records according to the methodology developed in sub-process 2.5 (Design processing and analysis). For example, weights can be used to "gross-up" data to make them representative of the target population (e.g. for sample surveys or extracts of scanner data), or to adjust for non-response in total enumerations. In other situations, variables may need weighting for normalisation purposes. It may also include weight correction for benchmarking indicators (e.g. known population totals).

5.7. Calculate aggregates

88. This sub-process creates aggregate data and population totals from microdata or lower-level aggregates. It includes summing data for records sharing certain characteristics (e.g. aggregation of data by demographic or geographic classifications), determining measures of average and dispersion, and applying weights from sub-process 5.6 (Calculate weights) to derive appropriate totals. In the case of statistical outputs which use sample surveys, sampling errors corresponding to relevant aggregates may also be calculated in this sub-process.

5.8. Finalise data files

89. This sub-process brings together the results of the other sub-processes in this phase in a data file (usually macro-data), which is used as the input to the "Analyse" phase. Sometimes this may be an intermediate rather than a final file, particularly for business processes where there are strong time pressures, and a requirement to produce both preliminary and final estimates.

Analyse Phase

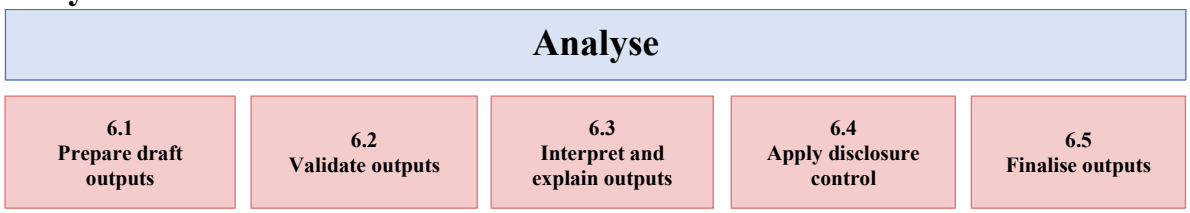

Figure 7. Analyse phase and its sub-processes

90. In this phase, statistical outputs are produced and examined in detail. It includes preparing statistical content (including commentary, technical notes, etc.), and ensuring outputs are “fit for purpose” prior to dissemination to users. This phase also includes the sub-processes and activities that enable statistical analysts to understand the data and the statistics produced. The outputs of this phase could also be used as an input to other sub-processes (e.g. analysis of new sources as input to the “Design” phase). For statistical outputs produced regularly, this phase occurs in every iteration. The "Analyse" phase and sub-processes are generic for all statistical outputs, regardless of how the data were sourced.

91. The "Analyse" phase is broken down into five sub-processes (Figure 7), which are generally sequential, from left to right, but can also occur in parallel, and can be iterative. These sub-processes are:

6.1. Prepare draft outputs

92. This sub-process is where the data from sub-processes 5.7 (Calculate aggregates) and 5.8 (Finalise data files) are transformed into statistical outputs such as indexes, seasonally adjusted statistics, e.g. trend, cycle, seasonal and irregular components, accessibility measures, etc., as well as the recording of quality characteristics such as coefficients of variation. The preparation of maps, GIS outputs and geo-statistical services can be included to maximise the value and capacity to analyse the statistical information.

93. Visualisation data on a map is useful for adding further geospatial context to data, but it also helps in detecting abnormalities during data validation (sub-process 6.2) and interpretation (sub-process 6.3).

94. When preparing the analysis output, it is important to pay attention to semantic interoperability, so that the output can be understood and used without ambiguities by users from different domains. Cataloguing and tagging the content using relevant metadata standards greatly increases the usability of the analysis outputs.

6.2. Validate outputs

95. This sub-process is where statisticians validate the quality of the outputs produced, in accordance with a general quality framework and with expectations. This sub-process includes activities involved with the gathering of intelligence, with the cumulative effect of building up a body of knowledge about a specific statistical domain. This knowledge is then applied to the current collection, in the current environment, to identify any divergence from expectations and to allow informed analyses. When outputs are designed and developed in partnership with external partners (e.g. NGIA), validation and quality requirements agreed with the partners should be checked. Validation activities can include:

- Checking that the population coverage and response rates are as required;
- Comparing the statistics with previous cycles (if applicable);
- Checking that the associated metadata, paradata and quality indicators are present and in line with expectations;
- Checking geospatial consistency of the data;
- Checking validity of geospatial information (e.g. boundary over contentious or disputed areas in the map);
- Confronting the statistics against other relevant data (both internal and external);
- Investigating inconsistencies in the statistics;
- Performing macro editing;
- Validating the statistics against expectations and domain intelligence.

6.3. Interpret and explain outputs

96. This sub-process is where the in-depth understanding of the outputs is gained by statisticians. They use that understanding to interpret and explain the statistics by assessing how well the statistics reflect their initial expectations, viewing the statistics from all perspectives using different tools and media, and carrying out in-depth statistical analyses such as time-series analysis, consistency and comparability analysis, revision analysis (analysis of the differences between preliminary and revised estimates), analysis of asymmetries (discrepancies in mirror statistics), geostatistical analysis using various GIS tools, etc.

6.4. Apply disclosure control

97. This sub-process ensures that the data (and metadata) to be disseminated do not breach the appropriate rules on confidentiality according to either organisation policies and rules, or to the process-specific methodology created in sub-process 2.5 (Design processing and analysis). This may include checks for primary and secondary disclosure, as well as the application of data suppression or perturbation techniques and output checking. The degree and method of statistical disclosure control may vary for different types of outputs. For example, the approach used for microdata sets for research purposes will be different to that for published tables, finalised outputs of geospatial statistics or visualisations on maps. Geospatially enabled statistics, in particular for high spatial resolution, may carry a greater risk of privacy breach and require in-depth disclosure control. When an interactive mapping application is used, the tool should be configured in a way that users are not allowed to drill down to the spatial resolution level that unit record data might be disclosed. Geospatial information adds a new dimension to data with which an individual statistical unit can be more easily identified in combination with other information. Therefore, a geospatial product component should be cross-checked with other components (e.g. tabular aggregates, anonymised micro datasets) before release so that it does not breach any privacy on its own as well as in combination with other outputs.

6.5. Finalise outputs

98. This sub-process ensures the statistics and associated information are fit for purpose and reach the required quality level and are thus ready for use. It includes:

- Completing consistency checks;
- Determining the level of release, and applying caveats;
- Collating supporting information, including interpretation, commentary, technical notes, briefings, measures of uncertainty and any other necessary metadata;
- Producing the supporting internal documents;

- Conducting pre-release discussion with appropriate internal subject matter experts;
- Translating the statistical outputs in countries with multilingual dissemination;
- Approving the statistical content for release.

Disseminate Phase

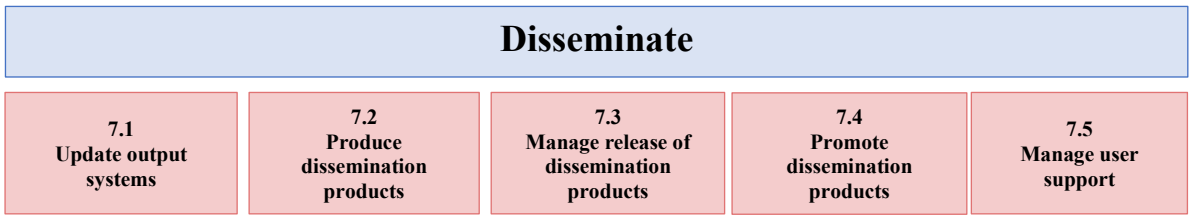

Figure 8. Disseminate phase and its sub-processes

99. This phase manages the release of the statistical products to users. It includes all activities associated with assembling and releasing a range of static and dynamic products via a range of channels. These activities support users to access and use the products released by the statistical organisation. For statistical products produced regularly, this phase occurs in each iteration.

100. Cataloguing and tagging the products using relevant metadata standards can greatly increase the discoverability and accessibility of the products both internally and externally. International standards should be used as a norm to ensure that the products can be found and consumed easily across a range of various user groups from public and private sector.

101. The "Disseminate" phase is broken down into five sub-processes (Figure 8), which are generally sequential, from left to right, but can also occur in parallel, and can be iterative. These sub-processes are:

7.1. Update output systems

102. This sub-process manages the update of systems (e.g. databases) where data and metadata are stored ready for dissemination purposes, including:

- Formatting data and metadata ready to be put into output systems;
- Loading data and metadata into output systems;
- Ensuring data are linked to the relevant metadata.

103. Formatting, loading and linking of metadata should preferably mostly take place in earlier phases, but this sub-process includes a final check that all of the necessary metadata are in place ready for dissemination.

7.2. Produce dissemination products

104. This sub-process produces the dissemination products, as previously designed in sub-process 2.1 (Design outputs), to meet user needs. They could include printed publications, press releases and websites. The products can take many forms including interactive graphics, tables, maps, public-use microdata sets, linked open data and downloadable files. Typical steps include:

- Preparing the product components (explanatory texts, tables, charts, maps, quality statements etc.);
- Assembling the components into products;
- Editing the products and checking that they meet publication standards.

105. When all product components are assembled together, additional disclosure control may be required at the product level (e.g. when a geospatial component such as map is provided by external partner organisations that has not undergone the disclosure control (sub-process 6.4)).

7.3. Manage release of dissemination products

106. This sub-process ensures that all elements for the release are in place including managing the timing of the release. It includes briefings for specific groups such as the press or ministers, as well as the arrangements for any pre-release embargoes. It also includes the provision of products to subscribers, and managing access to confidential data by authorised user groups, such as researchers. Sometimes an organisation may need to retract a product, for example, if an error is discovered. This is also included in this sub-process.

7.4. Promote dissemination products

107. Whilst marketing in general can be considered to be an overarching process, this sub-process concerns the active promotion of the statistical products produced in a specific statistical business process, to help them reach the widest possible audience. It includes the use of customer relationship management tools, to better target potential users of the products, as well as the use of tools including websites, wikis and blogs to facilitate the process of communicating statistical information to users. For a joint product, promotion and communication of the product may be coordinated with the partner organisation.

7.5. Manage user support

108. This sub-process ensures that user queries and requests for services such as microdata access are recorded, and that responses are provided within agreed deadlines. These queries and requests should be regularly reviewed to provide an input to the overarching quality management process, as they can indicate new or changing user needs. Replies to user requests can also be used to populate a knowledge database or a "Frequently Asked Questions" page, that is made publicly available, thus reducing the burden of replying to repeated and/or similar requests from external users. This sub-process also includes managing support to any partner organisations involved in disseminating the products. Geospatial information products could require additional support (e.g. tutorial video) as users may not be familiar with the data concept / type / structure, file format and associated GIS tools.

Evaluate Phase

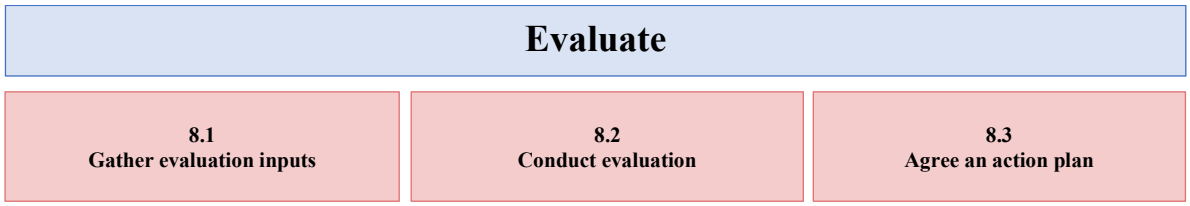

Figure 9. Evaluate phase and its sub-processes

109. This phase manages the evaluation of a specific instance of a statistical business process, as opposed to the more general overarching process of statistical quality management described in GSBPM Section VI (Overarching Processes). It can take place at the end of the instance of the process, but can also be done on an ongoing basis during the statistical production process. It relies on inputs gathered throughout the different phases. It includes evaluating the success of a specific instance of the statistical business process, drawing on a range of quantitative and qualitative inputs, and identifying and prioritising potential improvements.

110. It is important to note that the extent and depth of evaluation depend on the knowledge and experience of staff about the process under the evaluation. As geospatial production processes may be less well-established and understood, a systematic quality framework and capacity building of staff may be required for a thorough evaluation.

111. For statistical outputs produced regularly, evaluation should, at least in theory, occur for each iteration, determining whether future iterations should take place, and if so, whether any improvements should be implemented. However, in some cases, particularly for regular and well established statistical business processes, evaluation might not be formally carried out for each iteration. In such cases, this phase can be seen as providing the decision as to whether the next iteration should start from the "Specify Needs" phase, or from some later phase (often the "Collect" phase).

112. The "Evaluate" phase is broken down into three sub-processes (Figure 9), which are generally sequential, from left to right, but can also occur in parallel, and can be iterative. These sub-processes are:

8.1. Gather evaluation inputs

113. Evaluation material can be produced in any other phase or sub-process. It may take many forms, including feedback from users, process metadata (paradata), system metrics, and staff suggestions. Reports of progress against an action plan agreed during a previous iteration may also form an input to evaluations of subsequent iterations. This sub-process gathers all of these inputs, compiles quality indicators and makes them available for the person or team producing the evaluation. The collection of some of these evaluation materials can be automated and take place in a continuous way throughout the whole process, as defined by the quality framework (see Quality Management in GSBPM Section VI). On the other hand, for the evaluation of certain processes it can be necessary to perform specific activities such as small surveys, (e.g. post-enumeration surveys, re-interview studies, survey on effectiveness of dissemination).

8.2. Conduct evaluation

114. This sub-process analyses the evaluation inputs, compares them to the expected/target benchmarking results (when available), and synthesises them into an evaluation report or control dashboard. The evaluation can take place at the end of the whole process (ex-post evaluation) for selected activities, during its execution in a continuous way, or throughout the process, thus allowing for quick fixes or continuous improvement. The resulting report should note any quality issues specific to this iteration of the statistical business process as well as highlight any deviation of performance metrics from expected values, and should make recommendations for changes if appropriate. These recommendations can cover changes to any phase or sub-process for future iterations of the process, or can suggest that the process is not repeated.

8.3. Agree an action plan

115. This sub-process brings together the necessary decision-making power to form and agree an action plan based on the evaluation report. It should also include consideration of a mechanism for monitoring the impact of those actions, which may, in turn, provide an input to evaluations of future iterations of the process.

3. Overarching processes and corporate-level activities

116. The eight phases and the sub-processes within each phase of the GSBPM provide a set of building blocks that can be assembled in a sequence to create a production process. Section 2 describes the activities related to geospatial information and services that can be carried out in each phase to produce high-quality geospatially enabled statistics following the principles of the GSGF.

117. Some activities, however, are not limited to a certain phase of the process but rather apply throughout the entire production process (e.g. quality). Also, there are activities that should be conducted at a corporate level rather than as a part of specific production process as they support and influence many production processes across the organisation (e.g. management of address register, establishment of Memorandums of Understanding with other government agencies). In the GSBPM, activities of cross-cutting (across the production process) nature are modelled as "overarching process"[18] while activities at a corporate-level are covered by the Generic Activity Model for Statistical Organisation (GAMSO)[19], another HLG-MOS model complementing the GSBPM. Section 3 describes activities that are of cross-cutting nature and/or corporate-level.

Strategic collaboration and cooperation

118. Geospatial information are fundamental national information assets that can be used as a basis of numerous civic and commercial activities. A geospatial data ecosystem consists of various actors from both public (e.g. NGIA, transport department, agriculture agency, space agency) and private sector (e.g. utility companies, GIS service provider) providing and using a multitude of geospatial information and services. Therefore, coordination and cooperation within the ecosystem are critical to maximise synergy among different actors and avoid duplication of efforts. Depending on the regional and national context, coordination mechanism may vary (e.g. the INSPIRE directive of the European Union, the Sustainable Development Goal (SDG) data governance board in Ireland)[20]. Regardless of the format, active participation of statistical organisations in national geospatial information governance mechanisms, and continuous engagement with other agencies are important to ensure that standards, models and technologies are aligned as far as possible across the geospatial data ecosystem. This participation and engagement with the broader community also helps to ensure that the needs of statistical organisations are communicated and addressed in an efficient manner.

119. Geospatially enabled statistics with a high spatial resolution can provide invaluable input to other government agencies, in particular, to quickly respond to disasters and crises. Statistical organisations can play an important role in the geospatial data ecosystem not only as a producer of various statistical and geospatial information, but also as a provider of data integration services based on its extensive technical expertise. This service is becoming more and more essential in solving multi-faceted issues of the society such as climate change and migration. Statistical organisations can forge strategic collaborations with other agencies to produce high quality data integration products, which can demonstrate the vital function that the organisations can play in evidence-based decision-making processes with increasingly complicated challenges.

18 For more details about "overarching process", see GSBPM Section 2 and 4
19 For more information about GAMSO, see UNECE Statistics Wiki (https://statswiki.unece.org/display/GAMSO)
20 For more, see UN-GGIM "IGIF Strategic Pathway 1: Governance and Institutions" (https://ggim.un.org/IGIF/part2.cshtml)

Metadata management

120. Metadata provides essential information to understand and interpret the data (e.g. structure, classification used, analysis methodology, quality). It plays a key role to facilitate sharing, querying and discovery of data and services in an increasingly vast pool of data assets in statistical organisations. The use of metadata is not only limited to data products or services, but also includes various information that influences, triggers and regulates production processes (i.e. metadata-driven processes). The importance of metadata and its management is widely accepted in statistical organisations and much work has been done to develop standards and systems to manage metadata associated with typical statistical production process at the corporate level.

121. Compared to this statistical metadata, there is limited awareness and understanding in statistical organisations on the metadata associated with geospatial information and services. Geospatial information include various types ranging from orthoimagery (e.g. satellite data), elevation / depth, water to transport network[21], and this great variety of structures / formats as well as methodologies / technologies involved in producing the data adds challenges for statistical organisations to standardise geospatial metadata and systematically manage it.

122. Given the wide scope of geospatial metadata, it is important to first investigate crucial metadata elements (e.g. data type (point, line or polygon), time stamp, coordinate system) needed for different stages of production and determine a core metadata set and standards (e.g. ISO 19115, ISO 19119, GeoDCAT) to follow at the corporate level. After priorities are defined, a continuous improvement process could be put in place to gradually improve the scope covered by the metadata. It is important to have a corporate strategy in place to build a consistent metadata system to avoid compliance issues with existing metadata systems. As for statistical metadata, the geospatial metadata should be managed and updated continuously throughout the production process as the changes affect downstream tasks and influence the final outputs.

123. Alignment and harmonisation of geospatial metadata concepts with those of statistical metadata in existing metadata systems is critical, and there has been an increasing effort to connect statistical metadata with geospatial metadata (e.g. technical specification expanded for geospatial metadata in SDMX 3.0[22]). Emerging semantic web standards (e.g. Resource Description Framework (RDF) vocabularies) can provide the flexibility in modelling metadata as well as data[23] and its use for disseminating on the web offers a great potential to link outputs of statistical organisations with the plethora of data and resources on the web. Statistical organisations are encouraged to explore the semantic web standards as a long-term strategic objective with successive milestones to achieve dissemination of data and metadata within the framework of Linked Open Data (LOD).

---

[21] For more, see UN-GGIM "Global Fundamental Geospatial Data Themes" (http://ggim.un.org/documents/E-C20-2018-7-Add_1-Global-fundamental-geospatial-data-themes.pdf).

[22] Acknowledging the importance of geo-referencing statistical information, version 3.0 of SDMX will include technical specifications to improve the management of associated metadata. These technical specifications will help to connect the different levels of statistical information to geographical characteristics, making possible to include detailed geospatial structural and reference metadata in the exchange of statistical information. This work was done in collaboration with UN-GGIM, which is developing a Geospatial Roadmap to provide a bridge between statistical and geographical information.

[23] RDF-based ontologies such as Web Ontology Language (OWL) and Simple Knowledge Organization System (SKOS) can link resources using pre-defined properties such as owl:sameAs or skos: exactMatch.

Quality management

124. Quality is one of fundamental characteristics that define official statistics and its management throughout the production process has been a critical issue for statistical organisations. Quality is usually defined in terms of several dimensions on which various quality metrics are developed and agreed at the corporate level to ensure that quality is documented and monitored in a consistent and systematic manner for different processes across the organisation.

125. With growing use of administrative sources and the shift from single process / product to multiple processes / products, however, the management of quality has become increasingly challenging for statistical organisations. Examples of complications related to production of geospatially enabled statistics are grouped by input, processing and output aspects as below:

- (Input-aspect) For some types of geospatial information (e.g. earth observation data, network data), understanding the data quality often requires technical knowledge of the field[24];
- (Input-aspect) As statistical organisations move toward high-resolution products and point-based geocoding, the quality of geographies used in the production as an input has a greater impact on the process and the output, compared to when geographic units were at a coarser level (e.g. provincial, regional);
- (Input-aspect) While geospatial information may have quality terminologies similar to those in the field of statistics, they might be interpreted and calculated in a different way (e.g. term "precision" means spatial resolution rather than statistical variability[25]);
- (Processing-aspect) Geospatial information is often used as a basis for integrating data from different sources (e.g. survey, administrative data, big data). The quality of this integration is affected by the quality of geospatial information in each input dataset, but also greatly influenced by the quality of the geocoding or address-matching process itself;
- (Processing-aspect) More thorough disclosure control processing is needed as high-resolution geospatial information carries a greater risk of privacy infringement, not only on its own but also in combination with other data products;
- (Processing-aspect) There is a lack of understanding of the impact of geospatial information quality across different types of processing in the production process;
- (Output-aspect) Ensuring accessibility and usability of geospatial information and services could be challenging as there is a wide range of different requirements, priorities and needs depending on the user group. For example, researchers would require microdata equipped with geocodes readily available to integrate with other data sources for their analysis, city and municipal authorities would want datasets to be provided in formats that can be easily integrated within their local system, and journalists would be interested in key information and digestible headlines from geospatial analysis;
- (Output-aspect) With a widespread use of web-based map services, there is a greater expectation on the quality of map products from statistical organisations (e.g. visualisation, interactivity, user-friendliness).

[24] For example, see the comparison of quality concepts between statistical community and Earth Observation community in UN Global Working Group on Big Data - Earth Observation Data Task Team "UN Handbook on Satellite Data" (https://unstats.un.org/bigdata/task-teams/earth-observation/UNGWG_Satellite_Task_Team_Report_WhiteCover.pdf).

[25] H. Veregin (1999) "Data Quality Parameters"

126. Quality management that can be conducted as an overarching process or at a corporate level includes:

- Identifying the authoritative (external or internal) sources of reference data, establishing the quality profile of each data source based on the primary use cases within the organisation and communicating continuously with data providers regarding the quality requirements and associated risks;
- Establishing mechanisms with which feedback from production instances can be incorporated into the quality management processes of geospatial information holdings such as address registers (e.g. verification of geospatial information during field survey);
- Developing quality dimensions and metrics to be used at different stages of production and a consistent strategy at corporate level;
- Monitoring the new developments in the fast-evolving geospatial field and discussions at the regional and global levels to ensure that knowledge, methods and technologies in the organisations are up to date and in line with those of other communities.