High-level Workshop on Modernization of Official Statistics

Nizhni Novgorod, Russian Federation, 10-12 June 2014

Generic Statistical Business Process Model

Session 1: Standards and Tools for the Modernisation of Statistical Production and Services

Prepared by the High-Level Group for the Modernization of Statistical Production and Services

Summary

This document provides a description of the Generic Statistical Business Process Model and how it relates to other key standards for statistical modernisation. The Generic Statistical Business Process Model describes and defines the set of business processes needed to produce official statistics. It provides a standard framework and harmonised terminology to help statistical organisations to modernise their statistical production processes, as well as to share methods and components. The Generic Statistical Business Process Model can also be used for integrating data and metadata standards, as a template for process documentation, for harmonizing statistical computing infrastructures, and to provide a framework for process quality assessment and improvement.

Contents

Page

I. Introduction

1. The Generic Statistical Business Process Model (GSBPM) describes and defines the set of business processes needed to produce official statistics. It provides a standard framework and harmonised terminology to help statistical organisations to modernise their statistical production processes, as well as to share methods and components. GSBPM can also be used for integrating data and metadata standards, as a template for process documentation, for harmonizing statistical computing infrastructures, and to provide a framework for process quality assessment and improvement. These and other purposes for which GSBPM can be used are elaborated further in Section VI. This version of GSBPM is fully aligned with version 1.1 of the Generic Statistical Information Model (GSIM)[1] and provides a basis for the implementation of the Common Statistical Production Architecture (CSPA)[2].

A. Background

2. The Joint UNECE / Eurostat / OECD Work Sessions on Statistical Metadata (METIS) have prepared a Common Metadata Framework (CMF)[3]. Part C of this framework is entitled "Metadata and the Statistical Cycle". This part refers to the phases of the statistical business process (also known as the statistical value chain or statistical cycle) and provides generic terms to describe them. Since November 2013, this work has been taken over by the Modernisation Committee on Standards, under the High-Level Group for the Modernisation of Statistical Production and Services (HLG)[4].

3. During a workshop on the development of Part C of the CMF, held in Vienna in July 2007[5], the participants agreed that the business process model used by Statistics New Zealand would provide a good basis for developing a Generic Statistical Business Process Model. Following several drafts and public consultations, version 4.0 of GSBPM was released in April 2009. It was subsequently widely adopted by the global official statistics community, and formed one of the cornerstones of the HLG vision and strategy for standards-based modernisation.

4. In December 2012, a complementary model, the Generic Statistical Information Model (GSIM) was released (for more information see section III). The work to develop and subsequently implement GSIM resulted in the identification of several possible enhancements to GSBPM. During 2013, the HLG launched a project on "Frameworks and Standards for Statistical Modernisation" which included a broader review of GSBPM and GSIM, to improve consistency between the documentation of the models, and to incorporate feedback based on practical implementations. The current version of GSBPM (version 5.0) is the direct result of this work. Whilst it is considered final at the time of release, it is also expected that future updates may be necessary in the coming years, either to reflect further experiences from implementing the model in practice, or due to the evolution of the nature of statistical production. The reader is therefore invited to check the website www.unece.org/stats/gsbpm to be sure of having the latest version.

---

[1] See: http://www1.unece.org/stat/platform/display/metis/Generic+Statistical+Information+Model

[2] See: http://www1.unece.org/stat/platform/display/CSPA/Common+Statistical+Production+Architecture+Home

[3] See: http://www.unece.org/stats/cmf/

[4] See: http://www1.unece.org/stat/platform/display/hlgbas

[5] The papers from this Workshop are available at: http://www.unece.org/stats/documents/2007.07.metis.htm

B. Main changes from version 4.0

5. The review of version 4.0 of GSBPM had a clear mandate from the HLG to only introduce changes that had a strong business case and widespread support in the international statistical community. The rationale for this approach was that many organisations had invested heavily in implementing GSBPM. Major changes could imply significant costs, and could ultimately be counter-productive in terms of securing the widespread adoption and use of the model. As a result, there are a limited number of changes to the model itself, but many more improvements, additions and clarifications to the supporting documentation.

6. The main changes to GSBPM between versions 4.0 and 5.0 are as follows:

- Phase 8 (Archive) has been removed, and incorporated into the over-arching process of data and metadata management, to reflect the view that archiving can happen at any stage in the statistical production process.
- A new sub-process: "Build or enhance dissemination components" has been added within the "Build" phase to reflect the growing importance of having a range of dissemination options.
- Several sub-processes have been re-named to improve clarity.
- The descriptions of the sub-processes have been updated and expanded where necessary. The terminology used has been changed to be less survey-centric, in recognition of the growing use of non-survey sources (administrative data, big data etc.).

II. The Model

A. Understanding the Generic Statistical Business Process Model

7. GSBPM should be applied and interpreted flexibly. It is not a rigid framework in which all steps must be followed in a strict order, instead it identifies the possible steps in the statistical business process, and the inter-dependencies between them.

8. Although the presentation of GSBPM follows the logical sequence of steps in most statistical business processes, the elements of the model may occur in different orders in different circumstances. Also, some sub processes will be revisited a number of times forming iterative loops, particularly within the Process and Analyse phases.

9. GSBPM should therefore be seen more as a matrix, through which there are many possible paths. In this way GSBPM aims to be sufficiently generic to be widely applicable, and to encourage a standard view of the statistical business process, without becoming either too restrictive or too abstract and theoretical.

B. The structure

10. GSBPM comprises three levels:

- Level 0, the statistical business process;
- Level 1, the eight phases of the statistical business process;
- Level 2, the sub-processes within each phase.

11. A diagram showing the phases (level 1) and sub-processes (level 2) is included in Section IV. The sub-processes are described in detail in Section V.

12. GSBPM also recognises several over-arching processes that apply throughout the eight phases, and across statistical business processes. These can be grouped into two categories, those that have a statistical component, and those that are more general, and could apply to any sort of organisation. The first group ise considered to be more important in the context of this model, however the second group should also be recognised as they have (often indirect) impacts on several parts of the model.

13. Over-arching processes with a statistical component include the following. The first four are most closely related to the model. Quality and Metadata management are specifically highlighted in model diagrams and are elaborated further in Section VI.

- Quality management - This process includes quality assessment and control mechanisms. It recognises the importance of evaluation and feedback throughout the statistical business process;
- Metadata management - Metadata are generated and processed within each phase, there is, therefore, a strong requirement for a metadata management system to ensure the appropriate metadata retain their links with data throughout GSBPM. This includes process-independent considerations such as metadata custodianship and ownership, quality, archiving rules, preservation, retention and disposal;
- Data management - This includes process-independent considerations such as general data security, custodianship and ownership, data quality, archiving rules, preservation, retention and disposal;
- Process data management - This includes the management of data and metadata generated by and providing information on all parts of the statistical business process.
- Knowledge management - This ensures that statistical business processes are repeatable, mainly through the maintenance of process documentation;
- Statistical framework management - This includes developing standards, for example methodologies, concepts and classifications that apply across multiple processes;
- Statistical program management - This includes systematic monitoring and reviewing of emerging information requirements and emerging and changing data sources across all statistical domains. It may result in the definition of new statistical business processes or the redesign of existing ones;
- Provider management - This includes cross-process burden management, as well as topics such as profiling and management of contact information (and thus has particularly close links with statistical business processes that maintain registers);
- Customer management - This includes general marketing activities, promoting statistical literacy, and dealing with non-specific customer feedback.

14. More general over-arching processes include:

- Human resource management;
- Financial management;
- Project management;
- Legal framework management;
- Organisational framework management;
- Strategic planning.

C. Applicability

15. GSBPM is intended to apply to all activities undertaken by producers of official statistics, at both the national and international levels, which result in data outputs. It is designed to be independent of the data source, so it can be used for the description and quality assessment of processes based on surveys, censuses, administrative records, and other non-statistical or mixed sources.

16. Whilst the typical statistical business process includes the collection and processing of data to produce statistical outputs, GSBPM also applies to cases where existing data are revised or time-series are re-calculated, either as a result of improved source data, or a change in methodology. In these cases, the input data are the previously published statistics, which are then processed and analysed to produce revised outputs. In such cases, it is likely that several sub-processes and possibly some phases (particularly the early ones) would be omitted. Similarly, GSBPM can be applied to processes such as the compilation of National Accounts, and the typical processes of international statistical organisations

17. As well as being applicable for processes which result in statistics, GSBPM can also be applied to the development and maintenance of statistical registers, where the inputs are similar to those for statistical production (though typically with a greater focus on administrative data), and the outputs are typically frames or other data extractions, which are then used as inputs to other processes[6].

18. GSBPM should be seen as sufficiently flexible to apply in all of the above scenarios.

D. Using the Generic Statistical Business Process Model

19. GSBPM is a reference model. It is intended that GSBPM may be used by organisations to different degrees. An organisation may choose to either implement GSBPM directly or use it as the basis for developing an organisation specific adaption of the model. It may be used in some cases only as a model to which organisations refer when communicating internally or with other organisations to clarify discussion. The various scenarios for the use of GSBPM are all valid.

20. When organisations have developed organisation-specific adaptions of the GSBPM, they may make some specialisations to the model to fit their organisational context. The evidence so far suggests that these specialisations are not sufficiently generic to be included in the GSBPM itself.

21. In some cases it may be appropriate to group some of the elements of the model. For example, phases one to three could be considered to correspond to a single planning phase. In other cases, particularly for practical implementations, there may be a need to add one or more detailed levels to the structure presented below to separately identify different components of the sub-processes.

22. There may also be a requirement for a formal sign-off between phases, where the output from one phase is certified as suitable as input for the next. This sort of formal approval is implicit in the model, but may be implemented in many different ways depending on organisational requirements.

[6] See: http://www1.unece.org/stat/platform/download/attachments/57835551/BR+meeting+paper2.doc

III. Relationships with other models and standards

23. The GSBPM was originally developed based on the Generic Business Process Model from Statistics New Zealand, supplemented by input from other statistical organisations with experience of statistical process modelling. However, a number of other related models and standards exist, both as a result of modernisation initiatives, and similar activities outside the scope of official statistics.

A. Common Statistical Production Architecture

24. Turning first to the links between GSBPM and the other frameworks and standards needed for statistical modernisation, it can be helpful to consider them in the context of enterprise architecture. Enterprise architecture relates business functions and processes to the information, applications and technology needed to run them. It is a way of describing what an organisation does, and how it does it, to try to identify how the organisation could improve quality and efficiency.

25. In the context of statistical modernisation, the aim is to align the enterprise architectures of different organisations, creating an "industry architecture" for the whole "official statistics industry". This approach is intended to facilitate collaboration, sharing and joint development of the components and services that are needed for the different parts of the statistical business process (defined in relation to the GSBPM). The result is the Common Statistical Production Architecture (CSPA), first released at the end of 2013[7].

B. Generic Statistical Information Model

26. The Generic Statistical Information Model (GSIM) is a reference framework for statistical information, designed to play an important part in modernising and streamlining official statistics at both national and international levels. It enables generic descriptions of the definition, management and use of data and metadata throughout the statistical production process. It provides a set of standardised, consistently described information objects, which are the inputs and outputs in the design and production of statistics. GSIM helps to explain significant relationships among the entities involved in statistical production, and can be used to guide the development and use of consistent implementation standards or specifications.

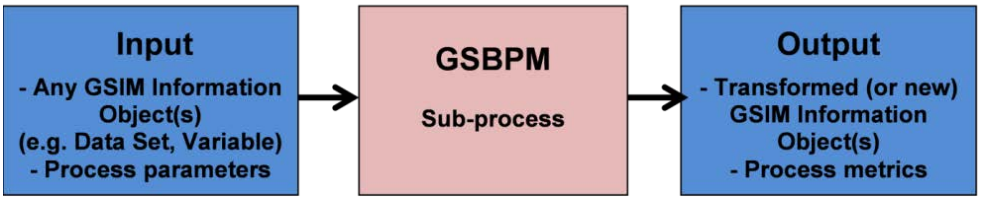
27. Like GSBPM, GSIM is one of the cornerstones for modernising official statistics and moving away from subject matter silos. GSIM is designed to allow for innovative approaches to statistical production to the greatest extent possible; for example, in the area of dissemination, where demands for agility and innovation are increasing. It also supports current approaches of producing statistics.

28. GSIM identifies around 110 information objects, examples include data sets, variables, statistical classifications, units, populations as well as the rules and parameters needed for production processes to run (for example, data editing rules).

29. GSIM and GSBPM are complementary models for the production and management of statistical information. As shown in the diagram below, GSIM helps describe GSBPM sub-processes by defining the information objects that flow between them, that are created

[7] See: http://www1.unece.org/stat/platform/display/CSPA/Common+Statistical+Production+Architecture+Home

in them, and that are used by them to produce official statistics. Inputs and outputs can be defined in terms of information objects, and are formalised in GSIM.



30. Greater value will be obtained from GSIM if it is applied in conjunction with GSBPM. Likewise, greater value will be obtained from GSBPM if it is applied in conjunction with GSIM. Nevertheless, it is possible (although not ideal) to apply one without the other. Similarly, both models support the implementation of CSPA, but can be applied regardless of whether that architectural framework is used or not.

31. In the same way that individual statistical business processes do not use all of the sub-processes described within GSBPM, it is very unlikely that all information objects in GSIM will be needed in any specific statistical business process.

32. Applying GSIM and GSBPM together can facilitate the building of efficient metadata driven systems, and help to harmonise statistical computing infrastructures.

C. Generic Longitudinal /Business Process Model

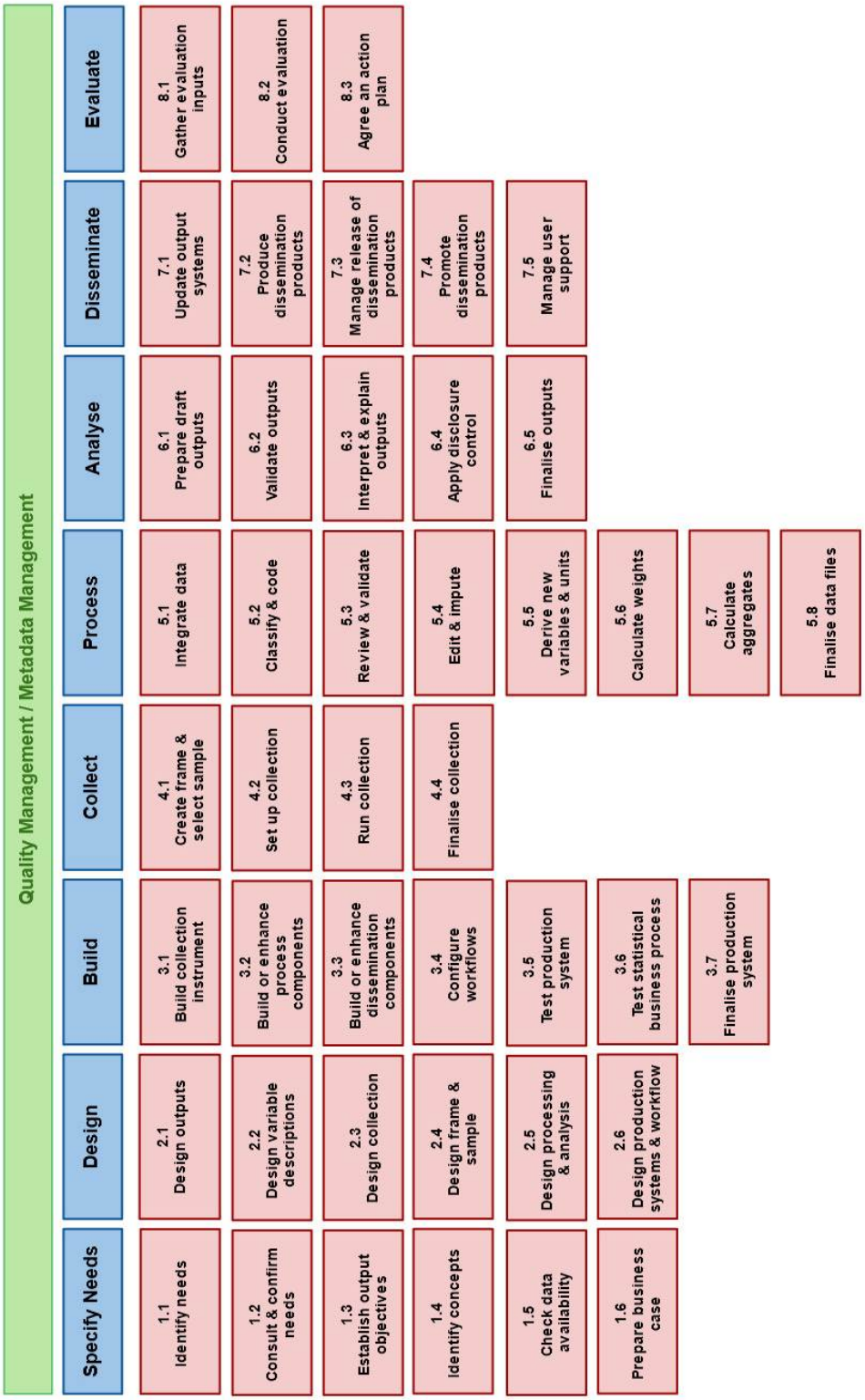
33. Looking outside the domain of official statistics, the social survey research community has developed the Generic Longitudinal Business Process Model (GLBPM)[8] "to provide a generic model that can serve as the basis for informing discussions across organisations conducting longitudinal data collections, and other data collections repeated across time".

34. Like GSBPM, GLBPM is a reference model against which actual business processes can be mapped. It is a reference model of the process of longitudinal and repeat cross-sectional data collection for research organisations. It describes the activities undertaken and maps these to their typical inputs and outputs.

35. GLBPM has many similarities to GSBPM, although it differs in some specific activities reflecting the different needs and practices of the statistical and social science research communities. GLBPM takes the approach of having a non-linear path through a matrix of alternatives directly from GSBPM.

[8] See: http://dx.doi.org/10.3886/DDILongitudinal05

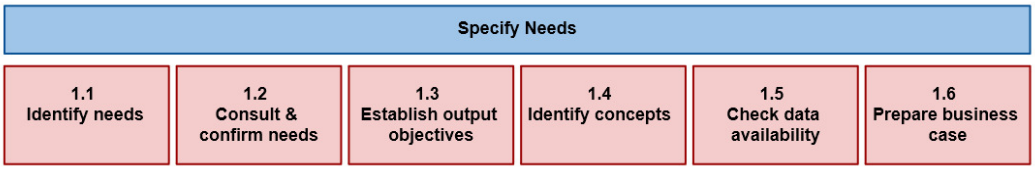
IV. Levels 1 and 2 of the Generic Statistical Business Process Model



V. Descriptions of phases and sub-processes

36. This section considers each phase in turn, identifying the various sub-processes within that phase, and describing their contents.

A. Specify Needs Phase



37. 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 customers to identify their detailed statistical needs, proposing high level solution options and preparing business cases to meet these needs.

38. In this phase the organisation:

- identifies the need for the statistics;
- confirms, in more detail, the statistical needs of the stakeholders;
- establishes the high level objectives of the statistical outputs;
- identifies the relevant concepts and variables for which data are required;
- checks the extent to which current data sources can meet these needs;
- prepares the business case to get approval to produce the statistics.

39. This phase is broken down into six sub-processes. These are generally sequential, from left to right, but can also occur in parallel, and can be iterative. The sub-processes are:

1. Identify needs

40. 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, 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 in particular the methods used by those organisations. It may involve consideration of specific needs of different user communities, such as the disabled, or different ethnic groups.

2. Consult and confirm needs

41. This sub-process focuses on consulting with the 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. For 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.

3. Establish output objectives

42. This sub-process identifies the statistical outputs 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. Legal frameworks (e.g. relating to confidentiality), and available resources are likely to be constraints when establishing output objectives.

4. Identify concepts

43. This sub-process clarifies the required concepts to be measured by the business process from the point of view of the user. At this stage the concepts identified may not align with existing statistical standards. This alignment, and the choice or definition of the statistical concepts and variables to be used, takes place in sub-process 2.2.

5. Check data availability

44. This sub-process checks whether current data sources 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 data sources, to determine whether they would be suitable for use for statistical purposes. When existing sources have been assessed, a strategy for filling any remaining gaps in the data requirement is prepared. 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.

6. Prepare business case

45. 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.

B. Design Phase

46. This phase describes the development and design activities, and any associated practical research work needed to define the statistical outputs, concepts, methodologies,

collection instruments[9] 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 statistical 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.

47. Design activities make substantial use of international and national standards, in order to reduce the length and cost of the design process, and enhance to comparability and usability of outputs. Organisations are also encouraged to reuse or adapt design elements from existing processes. Additionally, outputs of design processes may form the basis for future standards at the organisation, national or international levels.

48. This phase is broken down into six sub-processes, which are generally sequential, from left to right, but can also occur in parallel, and can be iterative. These sub-processes are:

1. Design outputs

49. 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. Disclosure control methods, as well as processes governing access to any confidential outputs are also designed here. Outputs should be designed to follow existing standards wherever possible, so inputs to this process may include metadata from similar or previous collections, international standards, and information about practices in other statistical organisations from sub-process 1.1 (Identify needs).

2. Design variable descriptions

50. This sub-process defines the statistical 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 classifications that will be used. It is expected that existing national and international standards will be followed wherever possible. 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 instrument may be inter-dependent to some degree. Preparation of metadata descriptions of collected and derived variables and classifications is a necessary precondition for subsequent phases.

3. Design collection

51. This sub-process determines the most appropriate collection method(s) and instrument(s). The actual activities in this sub-process will vary according to the type of collection instruments required, which can include computer assisted interviewing, paper questionnaires, administrative data interfaces and data integration techniques. This sub-process includes the design of 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 design of any formal agreements relating to data supply, such as memoranda of understanding, and confirmation of the legal basis for the data collection. 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

[9] For GSBPM purposes, collection instruments are defined broadly to include any tool or routine to gather or extract data and metadata, from paper questionnaires to web-scraping tools. In GSIM version 1.1, collection instruments are “exchange channels” used for incoming information.

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 process-specific provider management systems.

4. Design frame and sample

52. 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. 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.

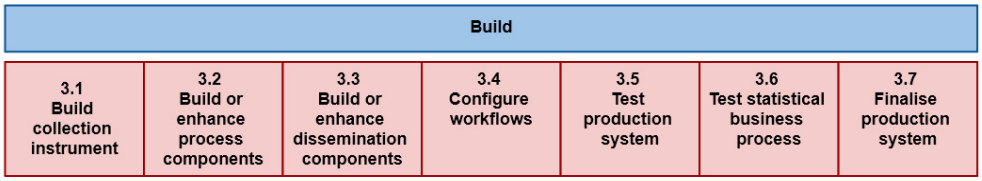
5. Design processing and analysis

53. This sub-process designs the statistical processing methodology to be applied during the "Process" and "Analyse" phases. This can include specification of routines for coding, editing, imputing, estimating, integrating, validating and finalizing data sets.

6. Design production systems and workflow

54. This sub-process determines the workflow from data collection to dissemination, taking an overview of all the processes required within the whole statistical production process, and ensuring that they fit together efficiently with no gaps or redundancies. Various systems and databases are needed throughout the process. 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 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.

C. Build Phase



55. 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 direct the selection of reusable processes, instruments, information, and services that 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 within the statistical production architecture.

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

57. It is broken down into seven sub-processes, which are generally sequential, from left to right, but can also occur in parallel, and can be iterative. These sub-processes are:

1. Build collection instrument

58. This sub-process describes the activities to build the collection instruments to be used during the "Collect" phase. The collection instrument is generated or 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 hubs. Collection instruments may also be data extraction routines used to gather data from existing statistical or administrative data sets. This sub-process also includes preparing and testing the contents and functioning of that instrument (e.g. testing the questions in a questionnaire). It is recommended to consider the direct connection of collection instruments to the statistical metadata system, so that metadata can be more easily captured in the collection phase. Connection of 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.

2. Build or enhance process components

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

3. Build or enhance dissemination components

60. This sub-process describes the activities to build new and enhance existing components and services needed for the dissemination of statistical products as designed in sub-process 2.1 (Design outputs). All types of dissemination components and services are included, from those that are used to produce traditional paper publications to those that provide web services, open data outputs, or access to micro-data.

4. Configure workflows

61. This sub-process configures the workflow, systems and transformations used within the statistical business processes, from data collection through to dissemination. It ensures that the workflow specified in sub-process 2.6 (Design production systems and workflow) works in practice.

5 Test production system

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 3.2 (Build or enhance process components), this sub-process also includes testing of interactions between assembled and configured services, and ensuring that the production solution works as a coherent set processes, information and services.

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 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

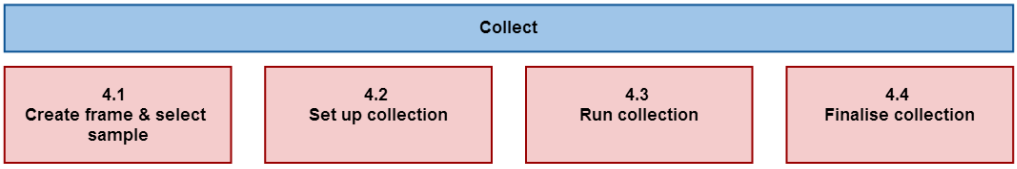
necessary to go back to a previous step and make adjustments to instruments, systems or components. For a major statistical business process, e.g. a population census, there may be several iterations until the process is working satisfactorily.

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 by business areas. The activities include:

- producing documentation about the process components, including technical documentation and user manuals
- training the business users on how to operate the process
- 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)).

D. Collect Phase



65. This phase collects or gathers all necessary information (data and metadata), using different collection modes (including extractions from statistical, administrative and other non-statistical registers and databases), 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.

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

1. Create frame and select sample

67. 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 statistical business process (for example to manage overlap or rotation), and between different processes using a common frame or register (for example 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 sources) as such processes generally create frames from the available data and then follow a census approach.

2. Set up collection

68. This sub-process ensures that the people, processes and technology 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. This sub-process includes:

- preparing a collection strategy;
- training collection staff;
- ensuring collection resources are available e.g. laptops;
- agreeing terms with any intermediate collection bodies, e.g. sub-contractors for computer assisted telephone interviewing
- 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 etc.).

69. For non-survey sources, this sub-process will include ensuring that the necessary processes, systems and confidentiality procedures are in place, to receive or extract the necessary information from the source.

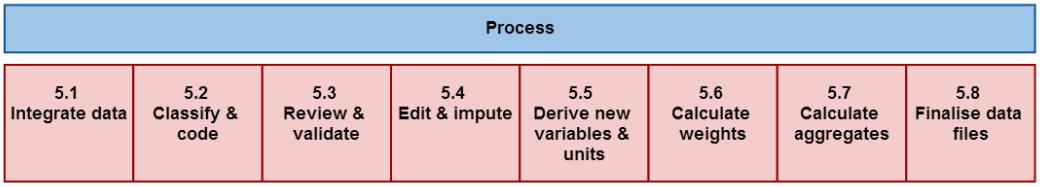
3. Run collection

70. This sub-process is where the collection is implemented, with the different instruments being used to collect or gather the information, which may include raw micro-data or aggregates produced at the source, as well as any associated metadata. It includes 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. 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. For administrative and other non-statistical sources, this process is brief: the provider is either contacted to send the information, or sends it as scheduled. When the collection meets its targets, it is closed and a report on the collection is produced. Some basic validation 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. All validation of the content takes place in the Process phase.

4. Finalise collection

71. 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 take-on, for example using clerical staff or optical character recognition tools to extract information from paper questionnaires, or converting the formats of files received from other organisations. It may also include analysis of the process metadata (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.

E. Process Phase



72. This phase describes the cleaning of data and their preparation for analysis. It is made up of sub-processes that check, clean, and transform input data, so that they can be analysed and disseminated as statistical outputs. It may be repeated several times if necessary. 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).

73. 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. Activities within the "Process" and "Analyse" phases may 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.

74. This phase is broken down into eight sub-processes, which may be sequential, from left to right, but can also occur in parallel, and can be iterative. These sub-processes are:

1. Integrate data

75. 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 data sources, and a variety of collection modes, including extracts of administrative 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
- matching / record linkage routines, with the aim of linking micro or macro data from different sources
- prioritising, when two or more sources contain data for the same variable, with potentially different values

76. 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 anonymised, that is stripped of identifiers such as name and address, to help to protect confidentiality.

2. Classify and code

77. 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 classification scheme.

3. Review and validate

78. This sub-process examines data to try 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 predefined 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. 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 web collection. Whilst this sub-process is concerned with detection of actual or potential errors, any correction activities that actually change the data are done in sub-process 5.4.

4. Edit and impute

79. Where data are considered incorrect, missing or unreliable, new values may be inserted 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:

- the determination of whether to add or change data;
- the selection of the method to be used;
- adding / changing data values;
- writing the new data values back to the data set, and flagging them as changed;
- the production of metadata on the editing and imputation process.

5. Derive new variables and units

80. 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.

6. Calculate weights

81. This sub process creates weights for unit data records according to the methodology created in sub-process 2.5 (Design processing and analysis). In the case of sample surveys, weights can be used to "gross-up" results to make them representative of the target population, or to adjust for non-response in total enumerations. In other situations, variables may need weighting for normalisation purposes.

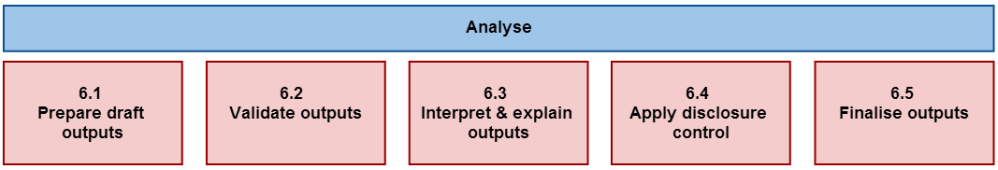
7. Calculate aggregates

82. This sub process creates aggregate data and population totals from micro-data or lower-level aggregates. It includes summing data for records sharing certain characteristics, determining measures of average and dispersion, and applying weights from sub-process 5.6 to derive appropriate totals. In the case of sample surveys, sampling errors may also be calculated in this sub-process, and associated to the relevant aggregates.

8. Finalise data files

83. This sub-process brings together the results of the other sub-processes in this phase and results in a data file (usually of 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.

F. Analyse Phase



84. In this phase, statistical outputs are produced, examined in detail and made ready for dissemination. It includes preparing statistical content (including commentary, technical notes, etc.), and ensuring outputs are “fit for purpose” prior to dissemination to customers. This phase also includes the sub-processes and activities that enable statistical analysts to understand the statistics produced. 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.

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

1. Prepare draft outputs

86. This sub-process is where the data are transformed into statistical outputs. It includes the production of additional measurements such as indices, trends or seasonally adjusted series, as well as the recording of quality characteristics.

2. Validate outputs

87. 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 also 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. 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 and paradata (process metadata) are present and in line with expectations
- 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.

3. Interpret and explain outputs

88. 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 produced for this cycle 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.

4. Apply disclosure control

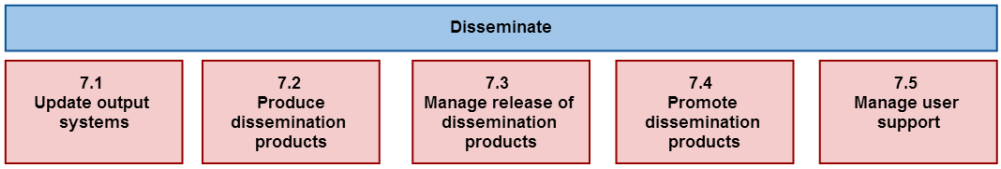
89. This sub-process ensures that the data (and metadata) to be disseminated do not breach the appropriate rules on confidentiality. This may include checks for primary and secondary disclosure, as well as the application of data suppression or perturbation techniques. The degree and method of disclosure control may vary for different types of outputs, for example the approach used for micro-data sets for research purposes will be different to that for published tables or maps.

5. Finalise outputs

90. 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;
- pre-release discussion with appropriate internal subject matter experts;
- approving the statistical content for release.

G. Disseminate Phase



91. This phase manages the release of the statistical products to customers. It includes all activities associated with assembling and releasing a range of static and dynamic products via a range of channels. These activities support customers to access and use the outputs released by the statistical organisation.

92. For statistical outputs produced regularly, this phase occurs in each iteration. It is made up of five sub-processes, which are generally sequential, from left to right, but can also occur in parallel, and can be iterative. These sub-processes are:

1. Update output systems

93. This sub-process manages the update of systems where data and metadata are stored ready for dissemination purposes, including:

- formatting data and metadata ready to be put into output databases;
- loading data and metadata into output databases;
- ensuring data are linked to the relevant metadata.

94. 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.

2. Produce dissemination products

95. This sub-process produces the products, as previously designed (in sub-process 2.1), to meet user needs. They could include printed publications, press releases and web sites. The products can take many forms including interactive graphics, tables, public-use micro-data sets and downloadable files. Typical steps include:

- preparing the product components (explanatory text, tables, charts, quality statements etc.);
- assembling the components into products;
- editing the products and checking that they meet publication standards.

3. Manage release of dissemination products

96. 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.

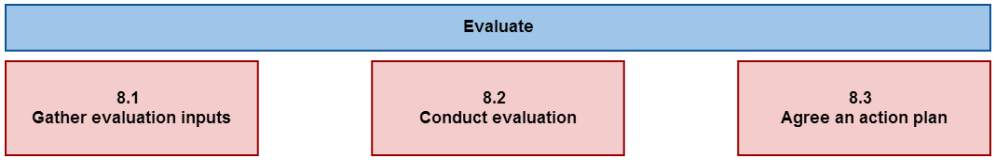
4. Promote dissemination products

97. Whilst marketing in general can be considered to be an over-arching 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 web sites, wikis and blogs to facilitate the process of communicating statistical information to users.

5. Manage user support

98. This sub-process ensures that customer queries and requests for services such as micro-data 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 over-arching quality management process, as they can indicate new or changing user needs.

H. Evaluate Phase



99. This phase manages the evaluation of a specific instance of a statistical business process, as opposed to the more general over-arching process of statistical quality management described in Section VI. It logically takes place at the end of the instance of the process, but 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.

100. 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 may 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).

101. This phase is made up of three sub-processes, which are generally sequential, from left to right, but which can overlap to some extent in practice. These sub-processes are:

1. Gather evaluation inputs

102. 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, and makes them available for the person or team producing the evaluation.

2. Conduct evaluation

103. This sub-process analyses the evaluation inputs and synthesises them into an evaluation report. The resulting report should note any quality issues specific to this iteration of the statistical business process, 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.

3. Agree an action plan

104. 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.

VI. Over-arching processes

105. GSBPM also recognises several over-arching processes that apply throughout the production phases, and across statistical business processes. Some of these over-arching processes are listed in paragraph 13 and 14. The processes of quality management and metadata management are further elaborated in this Section.

A. Quality management

106. Quality concerns organisations, processes and products. In the present framework, quality management over-arching process refers mainly to product and process quality.

107. The main goal of quality management within the statistical business process is to understand and manage the quality of the statistical products. There is general agreement among statistical organisations that quality should be defined according to the ISO 9000-2005 standard: "The degree to which a set of inherent characteristics fulfils requirements"[10]. Thus, product quality is a complex and multi-faceted concept, usually defined in terms of several quality dimensions. The dimensions of quality that are

[10] ISO 9000:2005, Quality management systems -- Fundamentals and vocabulary. International Organization for Standardization

considered most important depend on user perspectives, needs and priorities, which vary between processes and across groups of users.

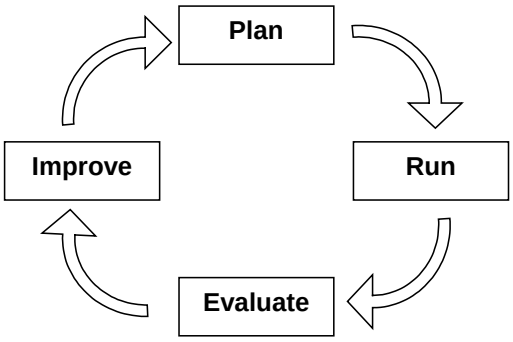
108. In order to improve the product quality, quality management should be present throughout the statistical business process model. It is closely linked to Phase 8 (Evaluate), which has the specific role of post-evaluating individual instances of a statistical business process. However, quality management has both a deeper and broader scope. As well as evaluating iterations of a process, it is also necessary to evaluate separate phases and sub-processes, ideally each time they are applied, but at least according to an agreed schedule. Metadata generated by the different sub-processes themselves are also of interest as an input for process quality management. These evaluations can apply within a specific process, or across several processes that use common components.

109. In addition, a fundamental role in quality management is played by the set of quality control actions that should be implemented within the sub-processes to prevent and monitor errors. The strategy could be reported in a quality assurance plan.

110. Within an organisation, quality management will usually refer to a specific quality framework, and may therefore take different forms and deliver different results within different organisations. The current multiplicity of quality frameworks enhances the importance of the benchmarking and peer review approaches to evaluation, and whilst these approaches are unlikely to be feasible for every iteration of every part of every statistical business process, they should be used in a systematic way according to a pre-determined schedule that allows for the review of all main parts of the process within a specified time period[11].

111. Broadening the field of application of the quality management over-arching process, evaluation of groups of statistical business processes can also be considered, in order to identify potential duplication or gaps.

112. All evaluations result in feedback, which should be used to improve the relevant process, phase or sub-process, creating a quality loop.



113. Examples of quality management activities include:

- Setting and maintaining of the quality framework;
- Setting of global quality criteria;
- Setting process quality targets and monitoring compliance;
- Seeking and analysing user feedback;

[11] A suitable global framework is the National Quality Assurance Framework developed by a global expert group under the United Nations Statistical Commission. See: http://unstats.un.org/unsd/dnss/QualityNQAF/nqaf.aspx

- Reviewing operation and documenting lessons learned;
- Examining process metadata and quality indicators;
- Internal or external auditing on statistical processes.

114. Quality management also involves institutional and organisational factors. Such factors are included in other GSBPM over-arching processes (e.g. Human resources management, Statistical programme management) although they can have an impact on quality.

B. Metadata management

115. Good metadata management is essential for the efficient operation of statistical business processes. Metadata are present in every phase, either created or carried forward from a previous phase. In the context of this model, the emphasis of the over-arching process of metadata management is on the creation, use and archiving of statistical metadata, though metadata on the different sub-processes themselves are also of interest, including as an input for quality management. The key challenge is to ensure that these metadata are captured as early as possible, and stored and transferred from phase to phase alongside the data they refer to. Metadata management strategy and systems are therefore vital to the operation of this model, and these can be facilitated by GSIM.

116. GSIM is a reference framework of information objects, which enables generic descriptions of the definition, management and use of data and metadata throughout the statistical production process. GSIM supports a consistent approach to metadata, facilitating the primary role for metadata envisaged in Part A of the Common Metadata Framework[12] "Statistical Metadata in a Corporate Context", that is, that metadata should uniquely and formally define the content and links between objects and processes in the statistical information system.

117. Part A of the Common Metadata Framework also identifies the following sixteen core principles for metadata management, all of which are intended to be covered in the over-arching Metadata Management process, and taken into the consideration when designing and implementing a statistical metadata system. The principles are presented in four groups:

1. Metadata handling

(a) *Statistical Business Process Model*: Manage metadata with a focus on the overall statistical business process model.

(b) *Active not passive*: Make metadata active to the greatest extent possible. Active metadata are metadata that drive other processes and actions. Treating metadata this way will ensure they are accurate and up-to-date.

(c) *Reuse*: Reuse metadata where possible for statistical integration as well as efficiency reasons

(d) *Versions*: Preserve history (old versions) of metadata.

2. Metadata Authority

(a) *Registration*: Ensure the registration process (workflow) associated with each metadata element is well documented so there is clear identification of ownership, approval status, date of operation, etc.

[12] See: http://www.unece.org/stats/cmf/PartA.html

(b) *Single source*: Ensure that a single, authoritative source ('registration authority') for each metadata element exists.

(c) *One entry/update*: Minimise errors by entering once and updating in one place.

(d) *Standards variations*: Ensure that variations from standards are tightly managed/approved, documented and visible.

3. Relationship to statistical cycle/process

(a) *Integrity*: Make metadata-related work an integral part of business processes across the organisation.

(b) *Matching metadata*: Ensure that metadata presented to the end-users match the metadata that drove the business process or were created during the process.

(c) *Describe flow*: Describe metadata flow with the statistical and business processes (alongside the data flow and business logic).

(d) *Capture at source*: Capture metadata at their source, preferably automatically as a by-product of other processes.

(e) *Exchange and use*: Exchange metadata and use them for informing both computer based processes and human interpretation. The infrastructure for exchange of data and associated metadata should be based on loosely coupled components, with a choice of standard exchange languages, such as XML.

VII. Other uses of the Generic Statistical Business Process Model

118. The original aim of GSBPM was to provide a basis for statistical organisations to agree on standard terminology to aid their discussions on developing statistical metadata systems and processes. However, as the model has developed, it has become increasingly apparent that it can be used for many other purposes, in particular related to modernisation of official statistics. A number of papers describing actual and potential uses of the GSBPM are available on the UNECE wiki platform[13]. The list below aims to highlight some current uses, and to inspire further ideas on how GSBPM can be used in practice.

- Harmonizing statistical production architectures - GSBPM can be seen as a model for an operational view of statistical computing architecture. It identifies the key components of the statistical business process, promotes standard terminology and standard ways of working across statistical business processes. It is a key enabler of the Common Statistical Production Architecture[14].
- Facilitating the sharing of statistical software - Linked to the point above, the GSBPM defines the components of statistical processes in a way that not only encourages the sharing of software tools between statistical business processes, but also facilitates sharing between different statistical organisations that apply the model. It has been used to "classify" software available for sharing in the inventory compiled by the Sharing Advisory Board[15].

[13] See: http://www1.unece.org/stat/platform/display/metis/Papers+about+the+GSBPM

[14] See: http://www1.unece.org/stat/platform/display/CSPA/Common+Statistical+Production+Architecture+Home

[15] See: http://www1.unece.org/stat/platform/display/msis/Software+Inventory

- Describing which standards are or could be used for different phases of the statistical production process. For example, Annex 2 of SDMX 2.1 User Guide[16] explores how SDMX applies to statistical work in the context of a business process model.
- Providing a framework for process quality assessment and improvement - If a benchmarking approach to process quality assessment is to be successful, it is necessary to standardise processes as much as possible. GSBPM provides a mechanism to facilitate this.
- Better integrating work on statistical metadata and quality - Linked to the previous point, the common framework provided by GSBPM can help to integrate international work on statistical metadata with that on data quality by providing a common framework and common terminology to describe the statistical business process.
- Providing the underlying model for methodological standards frameworks - Methodological standards can be linked to the phase(s) or sub-process(es) they relate to and can then be classified and stored in a structure based on GSBPM.
- Providing a structure for documentation of statistical processes - GSBPM can provide a structure for organizing and storing documentation within an organisation, promoting standardisation and the identification of good practices.
- Providing a framework for building organisational capability - GSBPM can be used to develop a framework assess the knowledge and capability that already exists within an organisation, and to identify the gaps that need to be filled to improve operational efficiency.
- Providing an input to high-level corporate work planning.
- Developing a business process model repository - Statistics New Zealand has developed a database to store process modelling outputs and allow them to be linked to their statistical business process model. They also plan to develop a Business Process Modelling Community of Practice - i.e. a regular forum to build knowledge of process modelling, to promote their business process model and increase understanding of it, and to discuss process modelling and models as enablers for process improvement.
- Measuring operational costs - GSBPM can be used as a basis for measuring the costs of different parts of the statistical business process. This helps to target modernisation activities to improve the efficiency of the parts of the process that are most costly.
- Measuring system performance - Related to the point above on costs, GSBPM can also be used to identify components that are not performing efficiently, that are duplicating each other unnecessarily, or that require replacing. Similarly it can identify gaps for which new components should be developed.

[16] See: http://sdmx.org/index.php?page_id=38

Annex 1– List of acronyms

Note: this list covers selected key abbreviations used in this paper. For a more comprehensive glossary of terms related to the statistical production process, see GSIM documentation.

CMF - Common Metadata Framework: A set of resources relating to the use of metadata by statistical organisations, including information on standards and best practices. See http://www.unece.org/stats/cmf/.

CSPA - Common Statistical Production Architecture: An industry architecture which brings together the GSBPM and GSIM, in addition to new frameworks about Statistical Services to create an agreed top level description of the 'system' of producing statistics which is in alignment with the modernisation initiative.

DDI - Data Documentation Initiative: An international standard for describing data from the social, behavioural, and economic sciences.

GLBPM - Generic Longitudinal /Business Process Model: A model based on the GSBPM, developed by the social survey research community.

GSBPM - Generic Statistical Business Process Model: A flexible tool to describe and define the set of business processes needed to produce official statistics.

GSIM - Generic Statistical Information Model: A reference framework of information objects, which enables generic descriptions of the definition, management and use of data and metadata throughout the statistical production process.

HLG - The High-Level Group for the Modernisation of Statistical Production and Services

METIS - The "brand name" for work on Statistical Metadata under the Conference of European Statisticians.

OECD - Organisation for Economic Cooperation and Development

SDMX - Statistical Data and Metadata eXchange: A set of technical standards and content-oriented guidelines, together with an IT architecture and tools, to be used for the efficient exchange and sharing of statistical data and metadata.

UNECE - United Nations Economic Commission for Europe

XML - eXtensible Mark-up Language: A language that defines a set of rules for encoding documents in a format that is both human-readable and machine-readable