Chapter 4a. The Generic Statistical Business Process Model

* 1. Introduction
	2. The model

## Introduction

The data collection, whether or not carried out through a survey, should not be considered in isolation but as part of a statistical process which takes into consideration various phases, from the identification of the users’ needs and objectives of the data collection, over design and processing steps to the final stages of data dissemination and evaluation. When planning a survey, for example, it is important to have a clear understanding of what the data needs are, what data are already available from other surveys or data sources and what type of information is intended to be disseminated and how.

Before describing different data collection methods for energy statistics in the next chapter of this publication, it is important to review the different stages of the statistical data production process and how they relate to each other. This chapter presents an overview of the Generic Statistical Business Process Model (GSBPM), which was developed by the United Nations Economic Commission for Europe (UNECE) to provide a standard framework for the business processes needed to produce official statistics. “The GBSPM can also be used for integrating data and metadata standards, as a template for process documentation, for harmonization statistical computing infrastructures, and to provide a framework for process data quality assessment and improvement.”[[1]](#footnote-1)

Because of its flexible nature, the GSBPM can of course be used to describe the collection and compilation of energy statistics. While the text of this chapter relies to a large extent to the description of the GSBPM available at XXXXXX, specific examples in energy statistics are provided throughout the chapter whenever relevant together with specific examples of issues to consider in the various stages of the GSBPM.

## The Generic Statistical Business Process Model

The GSBPM consists of 8 phases of the statistical business process and within each phase a number of sub-processes are further identified. Figure 4.1 provides an overview of the 8 phases and sub-processes of the GSBPM.

It should be noted that the GSBPM is not a rigid framework and should be applied with flexibility. The different phases of the GSBPM and their sub-processes do not have to be followed in the order described in the model and may be implemented in different order and sometimes in iterative ways. The various phases are described below.



Figure 1: The Generic Statistical Business Process Model

### Specify Needs Phase

This phase of the GSBPM is used to ascertain the need for the statistical information, establish output objectives based on the identified data needs; clarify the concepts from the point of view of the data users; check data availability and prepare a business vase in order to obtain an approval to carry out the new or modify statistical business model.

Identifying the users’ needs is particularly important as it defines the purpose for which the information is being collected and the uses this information could have in decision-making or research. It is good practice to consult potential data users at this stage and to elicit their guidance in developing the objectives and scope of the survey program. A clear statement of objectives, well-defined concepts and the level of data quality expected will allow the potential user to determine whether the information to be collected will serve their purposes. Consult with users to determine if the objectives align with their needs and whether there any other concepts or content to be considered. This will aid in developing a robust survey programme and help ensure the relevance of the statistical data being produced. Users can be organizations, groups, agencies and individuals who are expected to use the information for policy, research or other purposes.

Before designing a new survey process, there should be a review of the available data sources to determine whether information already exists to meet the users’ needs and the conditions under which that could be available. An assessment of possible alternatives would normally include research into potential administrative data sources and their methodologies, 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.

Once users have been consulted and their needs established, it is important to identify the conceptual, methodological and operational issues that should be addressed in the development of an energy statistical program. The preparation of a business case which includes also an assessment of costs and benefits associated with the undertaking the data collection project, the quality expected by users and the expected delivery dates of the survey data would be necessary to obtain the necessary approval for the overall data production process.

***Important issues to consider in this phase***

* When developing survey objectives and concepts, be sure to involve important users and other stakeholders.
* When developing a survey program, try to make it as cost-effective as possible. You may have to address the trade-off between cost and quality.
* Where explicit quality targets are established, ensure that they are included in the planning process.
* Elicit feedback from the group being surveyed (e.g. respondents in the energy sector) to test concepts and questions, to manage response burden and to increase respondent participation once the program goes into the collection stage.

###

### Design Phase

This phase of the GSBPM describes the development and design activities, and any associated practical research work needed to define the statistical outputs, concepts, methodologies, collection methods and operational processes. This phase also 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. During this phase the consistency with and the use the international and national standards are explicitly made in order to reduce the length and cost of the design process, and enhance to comparability and usability of outputs.

During this phase a number of sub-processes are identified. The identification of the content and quality of the final outputs is essential to be carried out at this stage of the data collection process as the subsequent work depends on the choice of the statistics to be disseminated and their format. Considerations for disclosure control methods should be also made explicit at this stage as well as the correspondence to existing standards.

Once the outputs have been properly specified and described, the relevant statistical variables to be collected can be identified and the data collection method and instrument can be determined. The actual activities in this sub-process vary according to the type of collection instruments required, which can include computer assisted interviewing, paper questionnaires, an electronic questionnaire, a telephone survey, administrative data interfaces and data integration techniques.

Designing a questionnaire should take into account the statistical requirements of data users and should be based on the outputs that were defined in the first stage. When designing a questionnaire there are a number of considerations important for an effective data collection; for example, the use of words and concepts that are easily understood by both respondents and questionnaire designers. It is usually the practice to consult with respondents during this stage about the content and wording to ensure their understanding of what is to be reported. Metadata requirements should also be specified (e.g. through the creation of a data dictionary). Ideally, data quality guidelines should also be established for questionnaire design.

If the chosen data collection method is based on sampling, the design of the frame and sample is decided in this phase of the GSBPM. The target population, that is the set of elements for which the estimated data are to represent, and the sampling frame should be clearly identified. Due the various constraints that exist in a statistical program, it is very likely that only a subset of this population will be used in compiling the statistics. This is the survey population or sample, the set of units that the constraints of the survey program necessitate us to narrow down to.

In the context of quality assurance, one desirable goal is to minimize the over-coverage or under-coverage between a target population and its sample. The frame should be kept as up-to-date as possible, as errors due to misclassification, duplication, erroneous conclusions or omissions will have a direct result on the quality of the estimates. Characteristics of the frame unit (e.g. contact information, address, size, identification) should also be of high quality as the these aspects of the frame are essential when used for stratification, collection, follow-up with respondents, data processing, imputation, estimation, record linkage, and quality assessment and analysis.

Finally, the workflow from collection to dissemination should be identified in order to ensure a smooth data collection process. Designing a workflow for a survey process should take into account the quality dimensions of accuracy, timeliness, credibility and cost-effectiveness.

Statistical processing methodology defines how the collected data will be treated after collection has closed. Designing a robust system is essential to ensuring the data quality of final estimates. The processing system should have clearly defined edit specifications that will be used to correct errors in the collected data and to help populate variables that are only partially reported. This ensures data quality by providing consistent treatment of partially completed surveys, and builds consistency checks into the initial phase of data capture and analysis.

The method for imputation for survey non-response should also be clearly defined. Determine which form of imputation will be used for which scenario and examine the implications of using a particular form of imputation. Some methods of imputation do not preserve the relationships between variables and may distort the underlying distributions. To ensure the highest possible quality, variance due to imputation (non-response error) should be taken into account when producing estimates.

Production systems are the means by which micro data are collected and processed into the final estimates. At this stage, it is essential to design a process flow that collects and processes the data through each stage of the survey process. Specifications should be designed that outline the needs of each stage of the survey process and that allow the output file of one stage to be loaded into the next.

***Important issues to consider in this phase***

Frame design:

* Test possible frames at the design stage of a survey for their suitability and quality.
* When more than one frame is available, consider using a combination of them if this would be expected to improve quality.
* Where possible, use the same frame or combination of frames when conducting multiple surveys with the same target population (e.g. for monthly and annual surveys).
* Be sure to actively monitor the frame quality on an ongoing basis by assessing its coverage of the desired target population and the accuracy of the description of the characteristics of the units (e.g. proper industrial coding, births/deaths).

Questionnaire design:

* When designing the questionnaire, use a flow and wording that allow the respondent to reply to questions as accurately as possible. Use clear and concise language.
* In the introduction to the questionnaire, include a survey title and subject and clearly identify the purpose of the survey. Indicate the authority under which the survey is being taken.
* To reduce errors made by the respondent, provide instructions that are clear, concise and noticeable.

Designing an imputation method:

* Develop and test imputation methods before implementation.
* When designing a survey program to produce energy statistics, take into account the relevant international standards in concepts and measurement procedures used in the collection and compilation of energy statistics as this will ensure their comparability.

###

### Build Phase

This phase of the GSBPM includes activities related to the building and testing of the data collection instrument and the data dissemination components, and the configuration of the workflow. Building any survey program requires careful coordination between its different elements. Designing a questionnaire should take into account the desired outputs and should be coordinated with the building of the processing system. The processing system should reflect the desired workflow and each stage of the production process should flow into the next. To ensure data quality, there should be an emphasis on testing once the questionnaire and collection, processing and imputation applications have been created to ensure that each system is integrated and functions properly. This stage includes the following activities:

The collection instrument is generated or built based on the design specifications created in the previous phase. A collection may use one or more modes to receive the data (e.g. personal or telephone interviews; paper, electronic or web questionnaires). Collection instruments may also be data extraction routines used to gather data from existing statistical or administrative data sets. This sub-process includes preparing and testing the contents and functioning of that instrument (e.g. testing the questions in a questionnaire). If possible, there should be a direct link between collection instruments and the statistical metadata system, so that qualitative information from respondents can be captured during the collection phase. The connection between 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.

Building and testing process and dissemination components includes, among other things, the detailed description of the subsequent phases of the GSBPM of “Process”, “Analyse” and “Disseminate”. Once built the processing system should be tested for functionality to be sure that the collected data will be processed correctly throughout the entire process.

Designing a process workflow involves configuring the flow of the data through the systems and transformations within the statistical business processes, from data collection through to the archiving of the final statistical outputs. Typically this form of testing involves 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.

Qualitative testing of the questionnaire should be conducted with respondents in the target population. This testing can consist of focus groups or in-depth one-on-one interviews, and can include cognitive testing. These methods are used to test question wording, sequencing and format. Cognitive testing involves assessing the respondents’ thought processes as they respond to the survey and determining whether they understand the questions and are providing accurate results. Qualitative testing may also be used to determine questionnaire content through the evaluation and exploration of key concepts.

Following the pilot, it may be necessary to go back to a previous step and make adjustments to instruments, systems or components.

Once testing is completed, the production systems can be finalized.

***Important issues to consider in this phase***

* All aspects of the newly designed collection and processing systems should be carefully tested.
* Consider two or more phases of questionnaire testing. This will allow testing for any revisions made to the questionnaire during development.
* If surveys are conducted with personal interviews, it is good practice to hold debriefing sessions with interviewers after the questionnaire has been tested.
* Thorough testing reduces the risk of any errors occurring during the production process that would delay the processing of the survey estimates.
* Develop the quality measures that will be used in subsequent stages of the survey process.

### Collect phase

Data collection is any process that acquires data to fulfill a survey objective. During this stage, data are acquired using different collection modes (including extractions from administrative and statistical registers and databases), and uploaded into the appropriate data environment. The collection stage includes the following sub processes:

*Create frame and select sample.* For a survey sample to be properly selected, the frame must be maintained and as up-to-date as possible in order to avoid imputations for non-existent establishments, to ensure all new establishments are included, to ensure the use of proper weights, to avoid misclassification, etc. Once the frame has been established, the sample is selected based on the criteria determined in the design phase. Quality assurance, approval and maintenance of the frame and the selected sample are also undertaken at this stage. This would include the maintenance of underlying registers, from which frames for several statistical business processes may be drawn. Please note that the survey itself can be used to subsequently update and maintain the frame by using the information collected.

*Set up collection*. Setting up the collection stage ensures that the people, processes and technology are ready to collect data, in all modes as designed. This includes the planning and training necessary to conduct the survey, which are essential to data quality. Training interviewers to ask effective questions and to minimize poor or non-response will ensure that questionnaires are correctly filled out and response rates are maximized.

*Run collection*. Once these stages have been completed, the collection process is ready to be implemented, using the different instruments to collect the data. This includes the initial contact with respondents and any subsequent follow-up or reminder actions. The process must record when and how respondents were contacted, and whether they have responded. This activity also includes the management of the respondents involved in the current collection, ensuring that the relationship between the statistical organization and data providers remains positive, recording comments, and addressing questions and complaints. For administrative data, this process is brief: the provider is either contacted to send the data, or sends it as scheduled. When the collection meets its targets (usually based on response rates) the collection is closed and a report on the collection is produced.

*Finalise collection.* 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. It may also include analysis of the process metadata 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.

### Process phase

This stage describes the cleaning of data records and their preparation for analysis. It is made up of sub-processes that check, correct, and transform the collected data so that they can be analysed and disseminated as statistical output. These sub-processes may be repeated several times.

During this phase of the GSBPM important steps in the compilation of the collected data take place, such the integration of data from different sources takes place: 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 review and validation of data as well as the editing and imputation of missing data are parts of data compilation and they are a fundamental step in the finalization of the data files.

***Important issues to consider in this phase***

* When undertaking the editing process, ensure that all edits are applied consistently.
* Put in place a strategy for selective follow-up. It may be helpful to introduce a score function that concentrates resources on the important units or the most severe errors that require follow-up.
* It is good practice to limit the editing process for fixing errors beyond a certain point as the contribution of further editing to error reduction diminishes. While editing is an essential process, it should have a cut-off point so as not to delay other activities.
* Be sure to identify, analyse and correct for extreme values, if necessary, as these will have a detrimental impact on the survey estimates.
* Guidelines in other sections of the survey process (e.g. editing, dissemination, etc) are also applicable to administrative data.
* When possible, collaborate with the designers of administrative systems or to help in building statistical requirements.
* Investigate and compare the quality of the administrative data with other sources if possible, the concepts, definitions, frequency, timeliness, and procedures underlying the collection and processing by the administrative organization.
* During imputation, evaluate the type of non-response and apply the appropriate imputation method.
* Be sure to carefully develop and test any imputation approach before it is introduced into the survey process.
* Ensure that imputed records are flagged and clearly identify the type of imputation applied to individual records. Evaluate the degree and effects of imputation.
* Consider the degree and impact of imputation during the analysis phase of the survey process.
* Proper estimation must conform to the sampling design. To that end, incorporate sampling weights in the estimation process as necessary.
* As the quality of the non-response adjustment factors used in the weighting process depends on model assumptions, validate the chosen model using diagnostics to minimize non-response bias.
* Outliers often lead to unreliable estimates when variables are continuous. When validating outliers, check for extreme values and outliers caused by excessive weights.

### Analyse phase

This phase of the GSBPM is where the statistical output are produced, examined in detail and made ready for dissemination. This process is what enables statistical analysts to understand the statistics produced.

There are a number of sub-processes defined within this phase which clearly describe the various steps to finalise the output: they include the validation of the statistical outputs; the in depth statistical analyses to better understanding out the statistics produced; and the application of disclosure control to ensure that the disseminated data and metadata do not breach confidentiality rules.

***Important issues to consider in this phase***

* Data should be validated prior to release. Use a variety of tools and processes to analyse the data and flag issues, then take corrective action.
* Determine whether the data are “fit for use.” If any data quality issues present themselves, assess the overall impact on the reliability of the data.
* Determine the extent of data quality evaluation required for a program or a product. Factors to consider include the data uses and users, risk of errors, quality variation over the life of the survey program, and the cost associated with data evaluation and its impact on data quality.
* At the minimum, information on coverage errors, response and imputation rates and sampling errors for key variables should be collected and analyzed to assess the quality of the estimates.
* Take the necessary steps to ensure the confidentiality of the data to be released.

### Disseminate phase

The dissemination phase of the GSBPM entails the release of the statistical output to the public. 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. For statistical outputs produced regularly, this phase occurs in each iteration.

The dissemination consists not only of the activities that are naturally associated with the data dissemination such as data publishing, but also activities related to: (a) the promotion of the dissemination products - through for examples, marketing of the products to reach the widest audience; (b) the management of the user support which ensure the prompt response to customers’’ queries as well as their regular review to provide an input to the overall quality management process as they can indicate new or changing user needs.

***Important issues to consider in this phase***

* Be sure to verify that the released data are consistent with the source data obtained. In the case of regrouped data or derived variables, this means that others should be able to reproduce the same results from the same source data.
* If data are being disseminated via an electronic product, the output should be thoroughly tested to ensure that errors are not being introduced by the programming.
* Provide data quality indicators so that end users are able to assess the quality and limitations of the disseminated data.
* When releasing a dissemination product, be consistent in style and formatting.
* It is good practice to provide a contact person for each release of information for users who have questions or feedback related to the data release.

### Evaluate phase

This phase of the GSBPM entails the regular evaluation of a specific instance of the statistical business process. All statistical activities should be reviewed periodically to adapt to evolving needs, to maintain the relevance and utility of the survey program, to address quality issues or concerns, to implement cost savings or strategies to reduce respondent burden, etc. The end of the evaluation phase is generally the agreement on an action plan to improve future iteration of the process.

***Important issues to consider in this phase***

* Review ongoing statistical activities periodically and adapt to evolving needs to maintain the relevance and responsiveness of your survey program.
* Elicit feedback from the data users about the relevance and quality of the data being produced. Attempt to implement their suggestions if a significant impact on data quality is foreseeable.
* Prioritize the quality issues based on their impact on the aggregate data. It is unlikely that correcting all the issues related to quality is possible in the timeframe allotted, so it is good practice to focus on those issues that have the most impact.
* Use this opportunity to determine if there are processes in the survey program that can be made more efficient.
* Review any user feedback pertaining to the accessibility of the data. Identify any possible improvements that can be implemented to improve the accessibility of the data to the end user.
* Identify any issues with the data or metadata that hinder accessibility or the end user’s ability to analyse the data.
* If possible, compare survey practices with the best practices of other statistical organizations and attempt to implement quality protocols that will increase the comparability of the data with those produced by other statistical organizations.
* However, note that implementing significant changes can create breaks in time series data. This can affect data comparability over time. Therefore, you must balance the needs for changes with the needs for consistency.
1. (UNECE) The Generic Statistical Business Process Model GSBPM (Version 5.0, December 2013) available at: http://www1.unece.org/stat/platform/display/GSBPM/GSBPM+v5.0 [↑](#footnote-ref-1)