Joint UNECE/Eurostat/OECD Work Session on Statistical Metadata (METIS)

Generic Statistical Business Process Model

Version 2.0 – August 2008

Prepared by the UNECE Secretariat [1]

I. Background

1. The METIS group has, over the last few years, been preparing a Common Metadata Framework (CMF), Part C of which 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. The initial intention was for statistical organizations to agree on standard terminology to aid their discussions on developing statistical metadata systems. Whilst this intention still remains valid, the use of this model can also be envisaged in other contexts, such as to help harmonize statistical computing infrastructures, to facilitate the sharing of software components, and to provide a framework for process quality assessment.

2. During the workshop on Part C of the CMF, held in July 2007, the participants agreed that the model currently used by Statistics New Zealand, with the addition of 'Archive' and 'Evaluate' phases, would provide a good basis for developing a "Generic Statistical Business Process Model" (referred to as "the model" in the rest of this paper). Variants on the New Zealand approach are already used by several statistical offices, the terms are generic enough to be broadly applicable and most importantly, the sub-processes that comprise each phase are documented to three levels, providing a sufficient amount of detail to clarify what is meant by each term.

3. A first draft of the model was prepared by the UNECE Secretariat, using the work of Statistics New Zealand as a basis, and was presented to the METIS Work Session in Luxembourg in April 2008. There have been a number of comments on the first draft, and considerable interest in the use of such a model outside the METIS group. As a result, this second draft is being circulated for further comments, with the aim of producing a more final version for a METIS Workshop planned for 2009. The timetable is therefore that comments on this draft are requested by the end of September 2008, allowing the next version to be produced before the end of the year.

[1] Prepared by Steven Vale (steven.vale@unece.org)

II. The Model

4. The model is intended to apply to statistical production regardless of the data source (surveys, administrative records, data integration etc.), and can be divided into three main levels (four if we count the model itself as Level 0):

- Level 1, the eight phases of the statistical business process;
- Level 2, the sub-processes within each phase;
- Level 3, a description of those sub-processes and their likely components.

5. The model also encompasses two over-arching processes that apply throughout the eight phases:

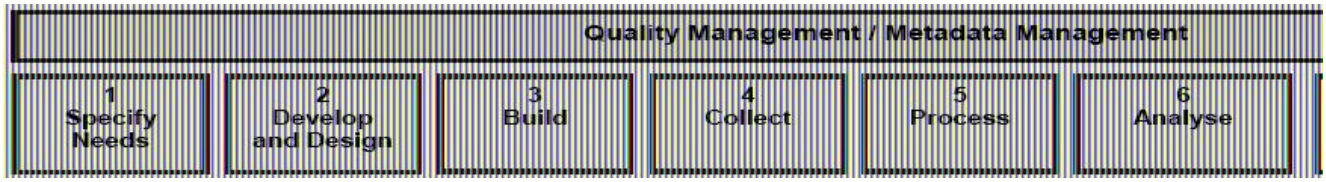
- Quality management – This process includes quality assessment and control mechanisms. It recognises the importance of evaluation and feedback throughout the statistical business process.
- Metadata management – Metadata are generated and processed within each phase, there is, therefore, a strong requirement for a metadata management system to ensure that the appropriate metadata retain their links with data throughout the model.

6. Other over-arching processes carried out in statistical agencies could be identified, but are deliberately not included in the model, as the model focuses exclusively statistical business processes. Administrative processes such as the management of human or financial resources are, by definition, excluded.

7. It is acknowledged that some statistical business processes will not use all phases of the model, and that some will only use certain phases or sub-processes for the first iteration, every n^{th} iteration, or following a change in methodology. Similarly, although the presentation follows the logical sequence of steps in most statistical business processes, the phases and sub-processes of the model may occur in different orders in different circumstances. The model should not, therefore be seen as a rigid framework, which has to be followed step-by-step. Instead it is a flexible tool to describe and define the set of business processes needed to produce official statistics.

8. The eight phases of level one are shown in the diagram below.

Level 1 of the Generic Statistical Business Process Model



A diagrammatic representation of the model, including all of the sub-processes (level 2) is included in Annex 1.

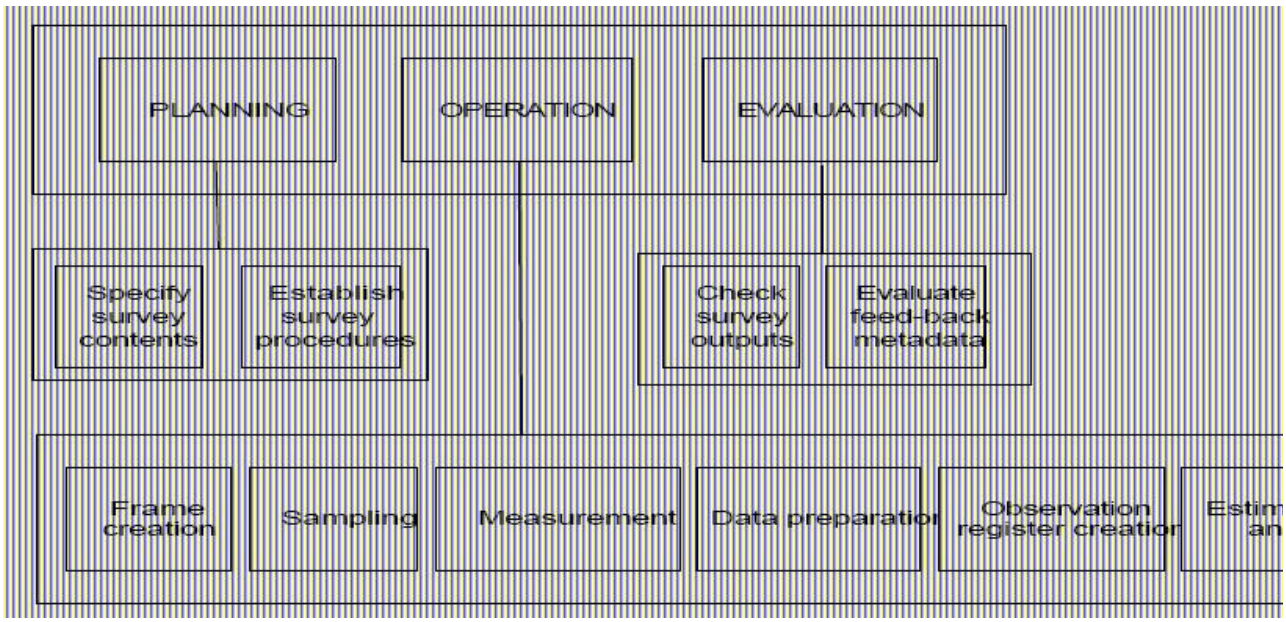
9. Each of the eight phases can be broken down into a number of sub-processes (level 2 of the model), each of which can be described in terms of their components (level3). These two levels are documented in detail in Annex 2, along with a more detailed description of the two over-arching processes.

III. Relationships with Other Models and Standards

10. The model has been developed drawing heavily on the Generic Business Process Model developed by Statistics New Zealand, as this is widely acknowledged an example of best practice in this area. However, a number of other related models and standards exist for different purposes and in different agencies, both at the national and international level. It would not be practical to give details of all national models here, but the main international models and standards are considered below, and related to the model proposed in this paper.

Information Systems Architecture for National and International Statistical Offices

11. This set of guidelines and recommendations was prepared by Professor Bo Sundgren and published by the United Nations in 1999. It contains the model of the phases and processes of a survey processing system shown below. Although different in presentation to the proposed generic model, the contents are largely the same.

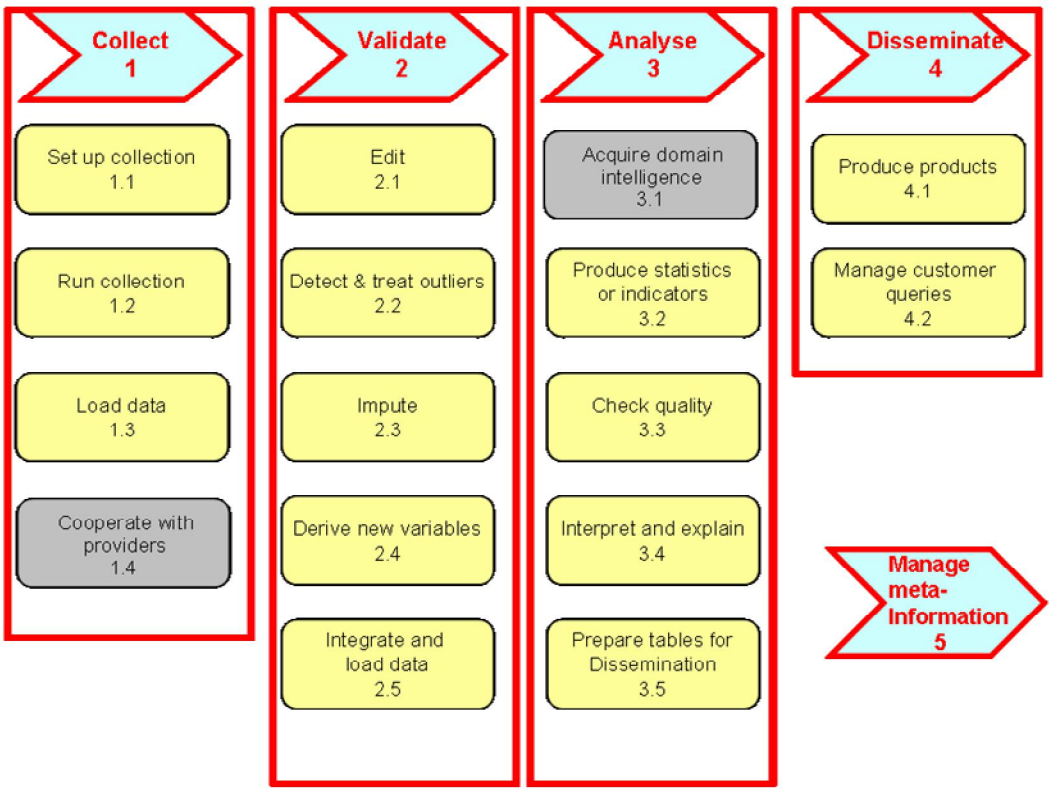


Source: Information Systems Architecture for National and International Statistical Offices – Guidelines and Recommendations, United Nations, 1999, http://www.unece.org/stats/documents/information_systems_architecture/1.e.pdf

The Eurostat "Cycle de Vie des Données" (CVD) model

12. This model was prepared as part of a major re-engineering of the statistical business process, and its components, within Eurostat. It fits well with phases four to

seven of the proposed generic model, and shows how this model can be adapted to the particular circumstances of an international statistical agency. The only significant difference is that “Manage meta-information” is treated as phase five in the CVD model, whereas it is covered within the over-arching process “Metadata Management” in the proposed generic model.



Source: Eurostat, Presentation of the CVD Implementation Plan, April 2008

The DDI 3.0 Combined Life Cycle Model

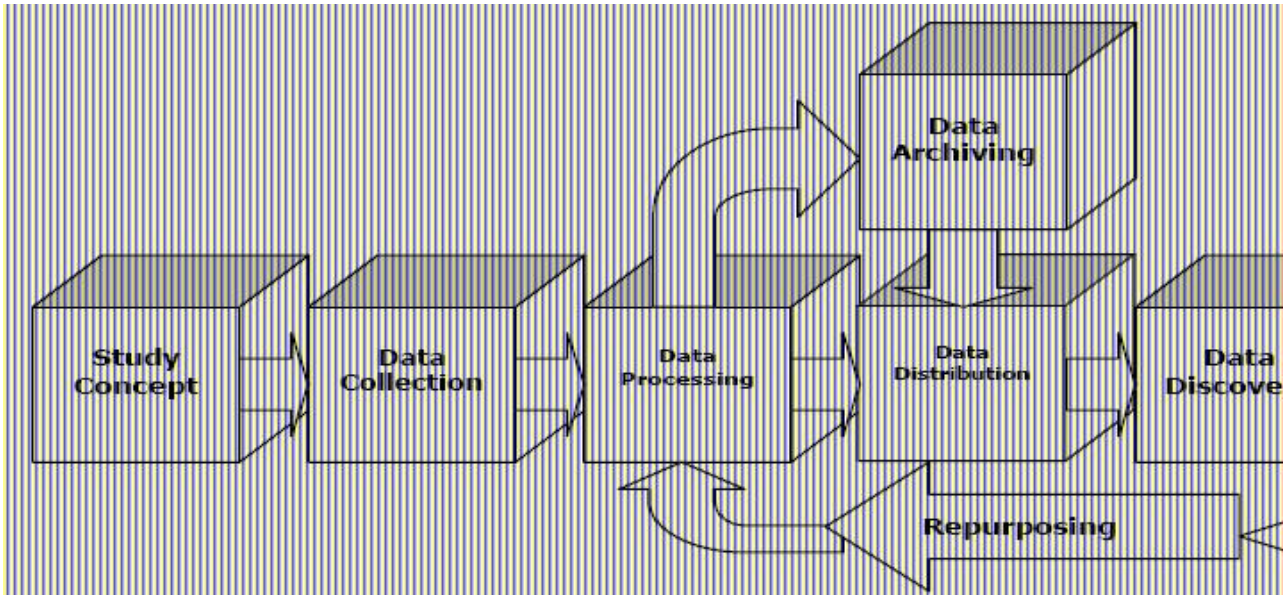
13. This model has been developed within the Data Documentation Initiative (DDI), an international effort to establish a standard for technical documentation describing social science data. The DDI Alliance comprises mainly academic and research institutions, hence the scope of the model below is rather different to the proposed generic model, which specifically applies to official statistical agencies. Despite this, the statistical business process appears to be quite similar between official and non-official statistics producers, as is clear from the high level of consistency between the models.

14. The main differences between the models are:

- The proposed generic model places data archiving at the end of the process, after the analysis phase.
- The DDI model replaces the dissemination phase with “Data Distribution” which takes place before the analysis phase. This reflects a difference in focus between the research and official statistics communities, with the latter putting a stronger emphasis on disseminating data, rather than research based on data disseminated by others.
- The DDI model contains the process of “Repurposing”, defined as the secondary use of a data set, or the creation of a real or virtual harmonized data set. This generally refers

to some re-use of a data-set that was not originally foreseen in the design and collect phases. This is covered in the proposed generic model in phase 1 (Specify Needs), where there is a sub-process to check the availability of existing data, and use them wherever possible. It is also reflected in the data integration sub-process within phase 5 (Process).

- The DDI model has separate phases for data discovery and data analysis, whereas these functions are combined within phase 6 (Analysis) in the proposed generic model.



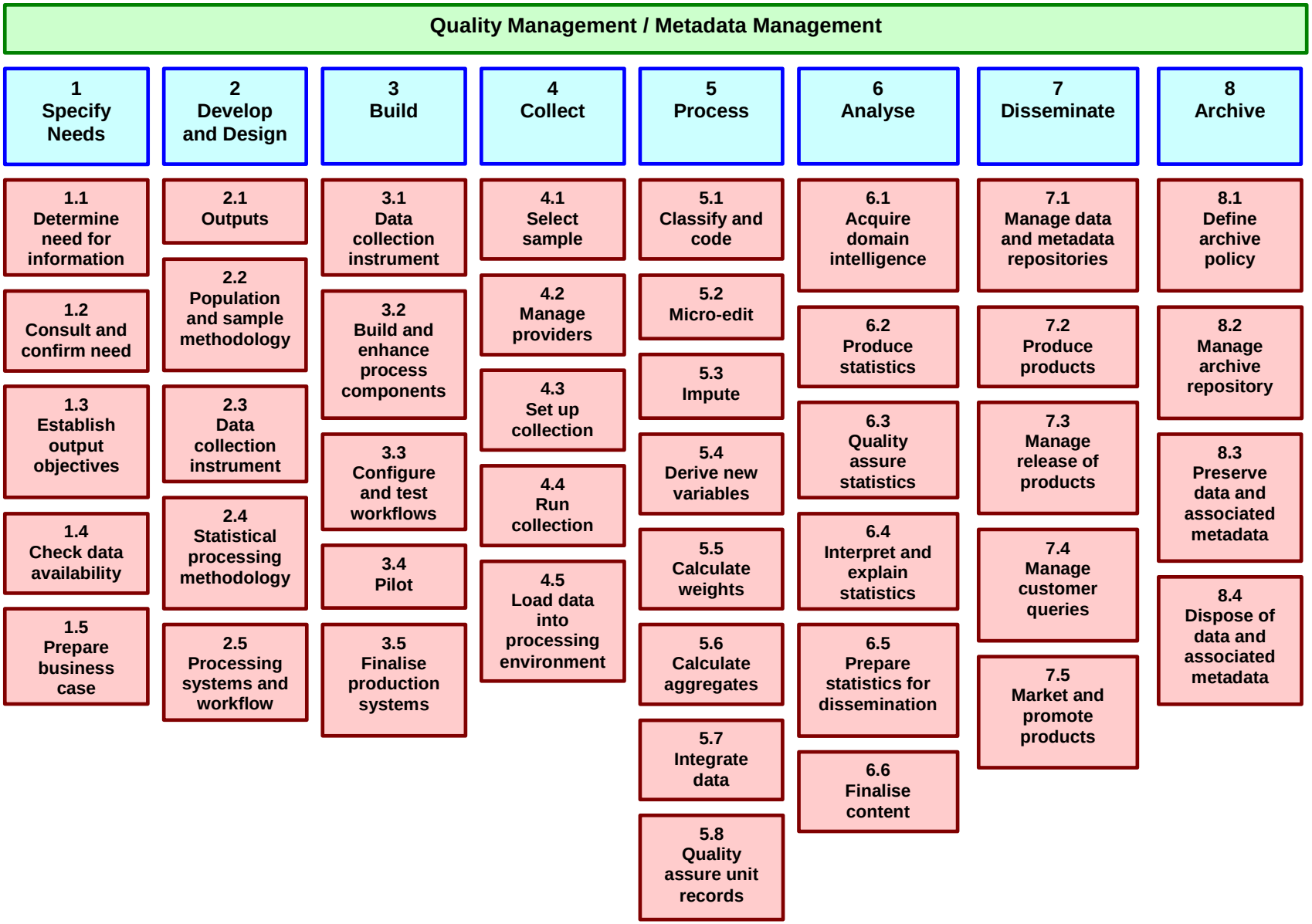
Source: Data Documentation Initiative (DDI) Technical Specification, Part I: Overview, Version 3.0, April 2008, http://www.ddialliance.org.

SDMX

15. The SDMX (Statistical Data and Metadata eXchange) set of standards do not provide a model for statistical business processes in the same sense as the three cases above. However they do provide standard terminology for statistical data and metadata, as well as technical standards for data and metadata transfer, which can be applied to transfers between sub-processes within a statistical agency.

16. The relationship between the model and SDMX was discussed at the April 2008 meeting of the METIS group. The final report of that meeting (paragraph 22) records a suggestion to incorporate the model into the Metadata Common Vocabulary and/or SDMX as a cross-domain concept. The model, in offering standard terminology for the different phases and sub-processes of the statistical business process, would seem to fit logically within the set of Content-oriented Guidelines developed for SDMX.

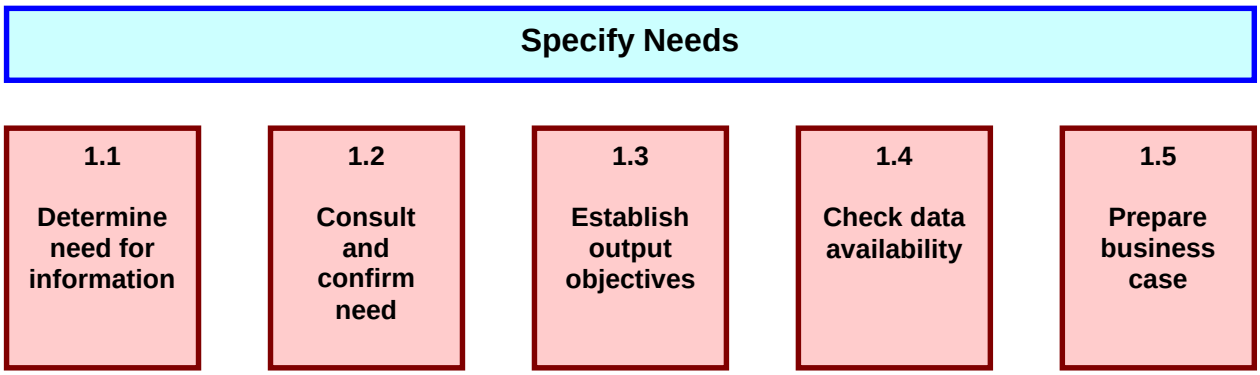
Annex 1 – Levels 1 and 2 of the Generic Statistical Business Process Model



Annex 2 – Levels 2 and 3 of the Generic Statistical Business Process Model

This annex considers each phase in turn, elaborating and describing the various sub-processes within that phase, and the components of those sub-processes. It therefore covers levels 2 and 3 of the model.

Phase 1 – Specify Needs



This phase determines whether there is a demand, externally and / or internally, for the identified statistics and whether the statistical agency can produce them. It is triggered when a need for new statistics is identified, or feedback about current statistics initiates a review.

In this phase the agency:

- determines the need for the statistics
- confirms, in more detail, the statistical needs of the stakeholders
- establishes the high level objectives of the statistical outputs
- checks if current collections and / or methodologies can meet these needs, and
- completes the business case to get approval to produce the statistics.

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

1.1. Determine need for information - This sub-process focuses on the initial research and broad identification of what statistics are needed and what is needed of the statistics. It also includes consideration of practice amongst other (national and international) statistical agencies producing similar data, and in particular the methods used by those agencies.

1.2. Consult and confirm need - This sub-process focuses on consulting with the stakeholders and confirming the in detail the need for the statistics. A good understanding of user needs is required so that the statistical agency 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.

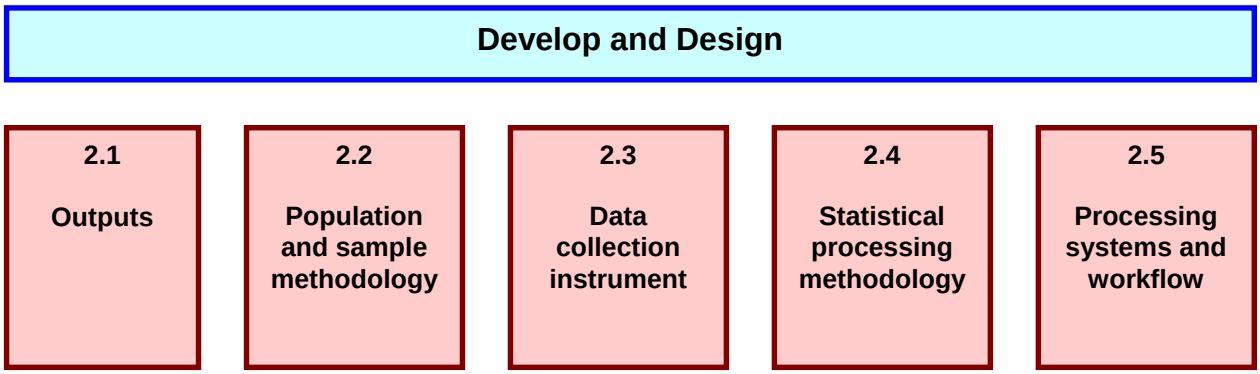
1.3. Establish output objectives - This sub-process focuses on identifying the statistical outputs that are required to meet the user needs identified in 1.2. It includes agreeing the suitability of the proposed outputs and their quality measures with users.

1.4. Check data availability - This sub-process focuses on identifying 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 also normally be included. When existing sources have been assessed, a strategy for filling any remaining gaps in the data requirement is prepared.

1.5. Prepare business case - This sub-process focuses on preparing a business case to get approval to implement the new or modified statistical business process. Apart from documentation from sub-processes 1.1 to 1.4, the contents of such a business case would typically include:

- 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.

Phase 2 – Develop and Design



This phase describes the research, development and design activities to define the statistical outputs, methodologies, collection instruments and operational processes. For statistical outputs produced on a regular basis, this phase usually occurs for the first iteration, and any subsequent reviews, rather than for every iteration.

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

2.1. Outputs – This sub-process focuses on the detailed design of the statistical outputs to be produced, including the related research and development work. Inputs may include metadata from similar or previous collections, international standards, and information about practices in other statistical agencies from sub-process 1.1.

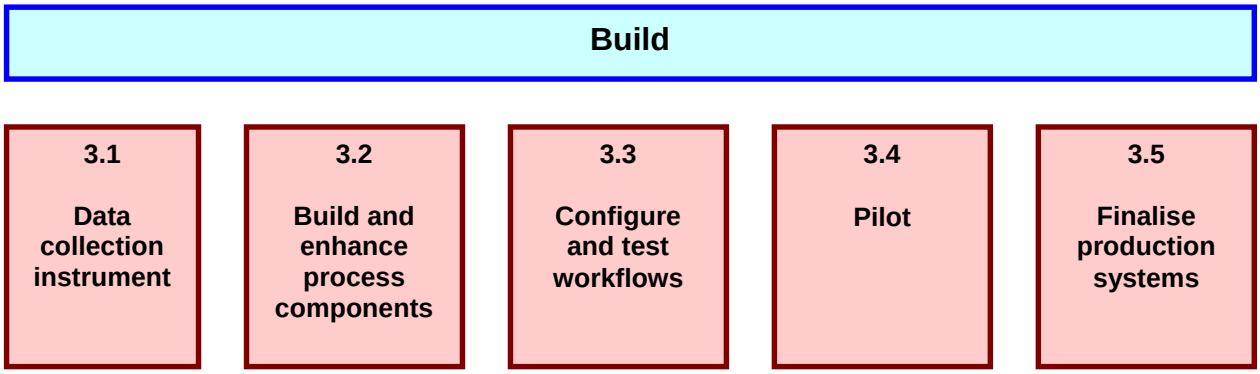
2.2. Population and sample methodology - This sub-process focuses on research, development and design work to identify and specify the population of interest, and to determine the most appropriate sampling methodology (which could include complete enumeration). For data integration projects, this step includes determining the population coverage of the linked data sets. The actual sample is created in Phase 4 - Collect (sub-process 4.1: Identify and validate sample), using the methodology, specified in this sub-process.

2.3. Data collection instrument - This sub-process focuses on research, development and design work to determine the most appropriate data collection instrument. The actual activities in this sub-process 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 is enabled by tools such as a question library, which facilitates the reuse of questions and related attributes, and a questionnaire tool, which enables the quick and easy compilation of questions into formats suitable for cognitive testing. The approved questions are used to build production ready collection instruments, regardless of collection mode, during the build phase (phase 3).

2.4. Statistical processing methodology - This sub-process focuses on researching, developing and designing the statistical processing methodology, to be applied during the process phase (phase 5). This can include developing and testing routines for coding, editing, imputing, estimating and integrating data.

2.5. Processing systems and workflow - This sub-process determines the workflow from data collection to archiving, taking an overview of all the processes required, and ensuring that they fit together efficiently with no gaps or redundancies. Existing components are examined to ensure they are fit for purpose for the collection in question, gaps are identified and solutions are proposed. A general principle is to reuse processes and technology across many statistical business processes.

Phase 3 – Build



This phase builds and tests the production systems to the point where they are ready for use in the “live” environment. 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, rather than for every iteration.

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

3.1. Data collection instrument - This sub-process describes the activities to build the collection instruments to be used during the Collect phase (phase 4). The collection instrument is generated or built based on the design specifications created during the Develop and Design phase. A collection may use one or more collection modes to receive the data, e.g. interviewers completing questions in person or over the telephone, or providers completing paper or web questionnaires. Collection instruments may also be data extraction routines used to gather data from existing statistical or administrative data sets. It also includes preparing and testing the contents and functioning of that instrument (e.g. testing the questions in a questionnaire).

3.2. Build and enhance process components - This sub-process describes the activities to build new and enhance existing software components needed for the business process, as designed in the “Develop and design” phase. Components may include dashboard functions and features, data repositories, transformation tools, workflow framework components and metadata management tools.

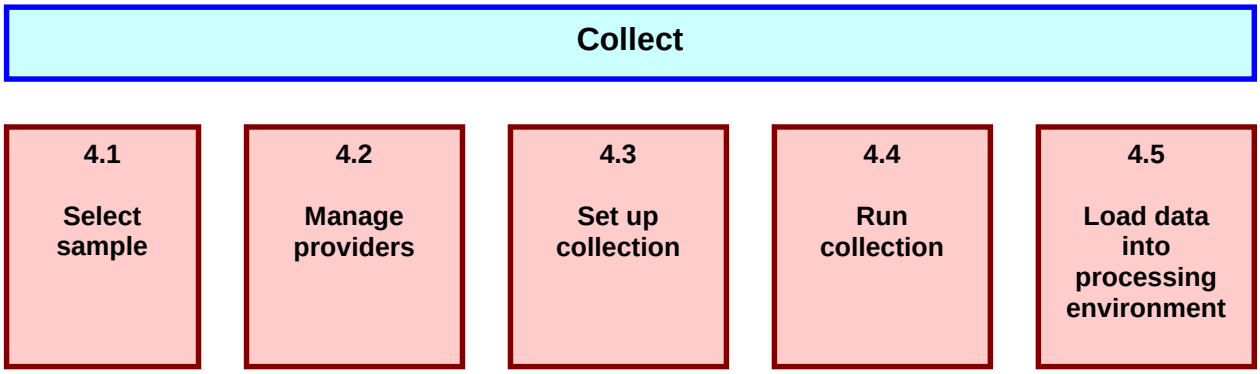
3.3. Configure and test workflow - This sub-process configures the workflow, systems and transformations used within the statistical business processes, from data collection, right through to archiving the final statistical outputs. It ensures that the workflow determined in 2.5 (Processing system and workflow) works in practice.

3.4. Pilot - 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.

3.5. Finalise production systems - This sub-process includes the activities to put the process, including workflow systems, modified and newly-built components 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 3.4, Pilot).

Phase 4 – Collect



This phase collects all external data, using different collection modes, and loads them into the appropriate data environment. For statistical outputs produced regularly, this phase occurs in each iteration. After the first iteration, it becomes part of “business as usual”.

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

4.1. Select sample - This sub-process creates the sample for this iteration of the collection. It also includes quality assurance and approval of the selected sample. It is not usually relevant for processes based entirely on the use of pre-existing data sources (e.g. administrative data), as such processes generally use all available data rather than a sample.

4.2. Manage providers - This sub-process is where the providers involved in current collections are managed. It takes place at anytime, during any period of the collection, and includes the management of:

- provider relationships, ensuring that the relationship between the statistical agency and data providers remains positive;
- provider load, managing and controlling the reporting burden on data providers;
- provider feedback, recording and responding to comments, queries and complaints.

4.3. Set up collection - This sub-process ensures that the people, processes and technology are ready to collect data, in all modes as designed. It takes place over a period of time, as it includes the strategy, planning and training activities for the collection. Where the collection is regular, these activities may not be explicitly required. For one-off and new surveys, these activities can be lengthy. This sub-process requires:

- an agreed collection strategy to be in place
- collection staff to be available and trained
- collection resources to be available e.g. laptops
- collection systems to be configured to request and receive the data, and
- collection data security to be ensured.

4.4. Run collection - This sub-process is where the collection is implemented, with the different collection instruments used to collect the data. 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.

4.5. Load data into processing environment - This sub-process includes initial data validation, as well as loading the data (and metadata) into a suitable electronic environment for further processing in phase 5. It may include automatic data take-on, for example using optical character recognition tools to extract data from paper questionnaires, or converting the formats of data files received from other organisations. In cases where there is a physical data collection instrument, such as a paper questionnaire, which is not needed for further processing, this sub-process manages the archiving of that material in conformance with the principles established in phase 8.

Phase 5 – Process

Process

5.1 Classify and code	5.2 Micro -edit	5.3 Impute	5.4 Derive new variables	5.5 Calculate weights	5.6 Calculate aggregates	5.7 Integrate data	5.8 Quality assure unit records

This phase describes the cleaning of data records and their preparation for analysis. It is made up of sub-processes that check, clean, and transform the collected data, and may be repeated several times. For statistical outputs produced regularly, this phase occurs in each iteration. After the first iteration, it becomes part of business as usual. The sub-processes in this phase can apply to data from both statistical and non-statistical sources (with the possible exception of sub-process 5.5, Calculate and apply weights, which is usually specific to survey data).

The “Process” and “Analyse” phases are 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.

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

5.1. Classify and code - This sub-process classifies and codes the input data. For example automatic (or clerical) coding routines may assign numeric codes to text responses. It is also where statistical units can be standardised, anonymized and assigned a unique code. Anonymization is where data are stripped of identifiers such as name and address, to help to protect confidentiality.

5.2. Micro-edit - The micro-edit sub-process applies to each record, and looks at the unit record data to try to identify and, where necessary, to correct errors and discrepancies. It may be run iteratively, validating data against predefined edit rules, usually in a set order. It may apply auto-edits, or raise alerts for manual inspection and correction of the data. Micro-editing applies to unit records from all types of collections, and can take place before and after data integration.

5.3. Impute - This sub-process identifies missing data, both at the variable level (item non-response), and at the unit level (unit non-response), and uses a rule-based approach to fill these gaps with estimated values. It includes:

- the selection of data to include or exclude from the imputation routine;

- imputation using one or more pre-defined methods e.g. “hot-deck”, “cold-deck” or other methods;
- writing the imputed data to the data set, and flagging them as imputed;
- the production of metadata on the imputation process;

Imputation applies to unit records both from surveys and administrative sources, before and after integration.

5.4. Derive new variables - This sub-process creates variables that are not explicitly provided in the collection and are needed to deliver the required outputs. It derives these new variables by applying arithmetic formulae to one or more of the variables that are already present in the dataset. It 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.

5.5. Calculate weights - This sub process creates weights for unit data records according to the methodology created in sub-process 2.4: Research, develop and design statistical processing methodology. These weights are generally used to “gross-up” sample survey results to make them representative of the target population.

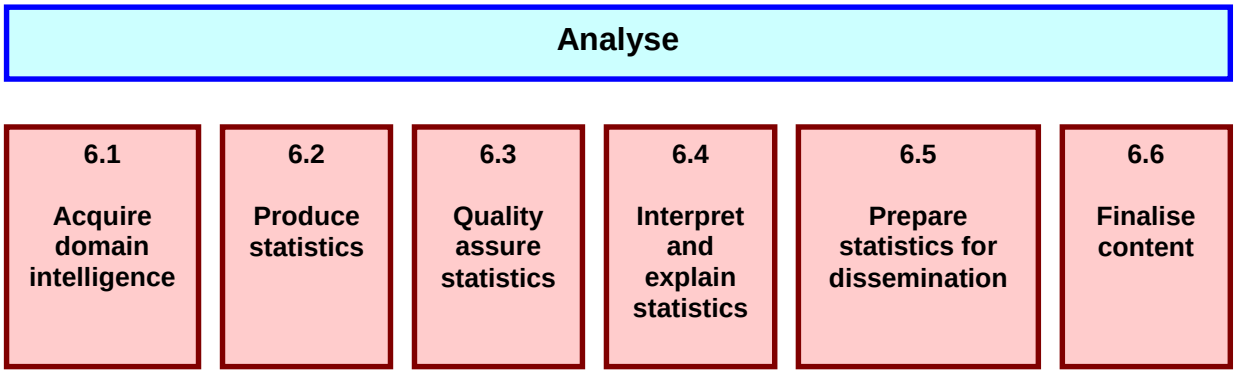
5.6. Calculate aggregates - This sub process creates aggregate data and population totals from micro-data. It includes summing data for records sharing certain characteristics, determining measures of average and dispersion, and applying weights from sub-process 5.5 to sample survey data to derive population totals.

5.7. Integrate data - This sub-process integrates one or more data sources. The input data can be from a mixture of external or internal data sources, and a variety of collection modes. The result is a harmonised data set. Data integration typically includes:

- matching / record linkage, to ensure data referring to the same units are correctly linked;
- prioritising when two or more sources contain the same variable (often with different values);
- resolving gaps and inconsistencies in the integrated data set (by invoking sub-processes 5.2, 5.3 and 5.4);
- quality assuring the results of the data integration sub-process.

5.8. Quality assure unit records - This sub-process ensures that the data set resulting from phase 5 is fit for use in subsequent phases.

Phase 6 – Analyse



In this phase, statistics are produced, examined in detail, interpreted, understood and made ready for dissemination. This phase 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. After the first iteration, it becomes part of business as usual. The Analyse phase and sub-processes are generic for all statistical outputs, regardless of how the data were sourced.

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

6.1. Acquire domain intelligence - This sub-process includes many ongoing 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 allow informed analyses. Acquiring a high level of domain intelligence will allow a statistical analyst to understand the data better, and to identify where results might differ from expected values. This allows better explanations of these results in sub-process 6.4.

6.2. Produce statistics - This sub-process is where domain intelligence is applied to the data collected to produce statistical outputs. It includes the production of additional measurements such as indices or seasonally adjusted series.

6.3. Quality assure statistics - This sub-process is where statisticians verify the quality of the statistics produced, in accordance with a general quality framework. Verification activities can include:

- checking that the population coverage and response rates are as required;
- comparing the statistics with previous cycles (if applicable);
- confronting the statistics against other relevant data (both internal and external);
- investigating irregular information in the statistics (e.g. outliers);
- performing macro editing;
- verifying the statistics against expectations and domain intelligence.

6.4. Interpret and explain statistics - This sub-process is where the in-depth understanding of the statistics 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.

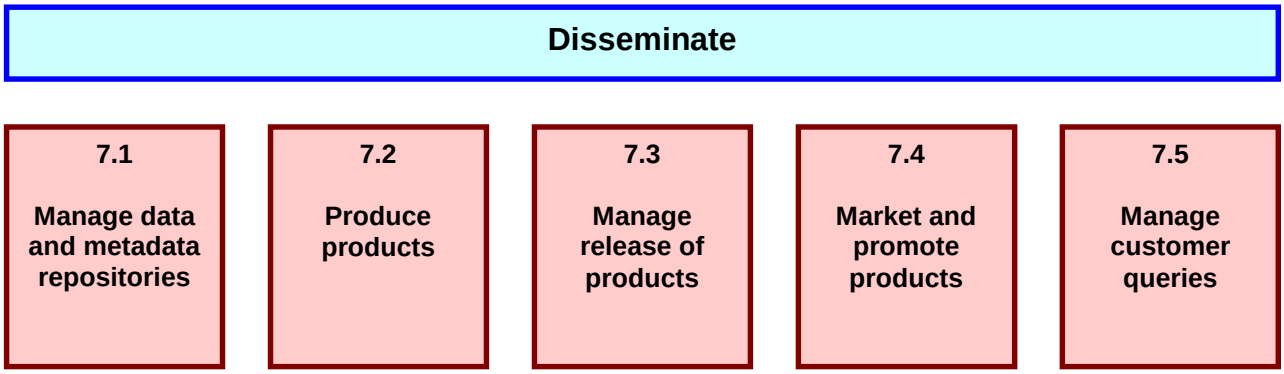
6.5. Prepare statistics for dissemination - This sub-process ensures the statistics and associated information are ready for dissemination by:

- applying data confidentiality rules;
- determining the level of release, and applying caveats;
- producing the supporting information, including any interpretation, and all necessary metadata to accompany the statistics;
- producing the supporting internal documents.

6.6. Finalise content - This sub-process ensures that the statistics and corresponding documentation to be released are fit for purpose and reach the required quality level. The activities include:

- pre-release discussion with related internal subject matter experts;
- finalising the information and explanation;
- completing consistency checks;
- approving the statistical content for release.

Phase 7 – Disseminate



This phase manages the release of the statistical products to customers. For statistical outputs produced regularly, this phase occurs in each iteration. After the first iteration it becomes part of business as usual. This phase is made up of five sub-processes, which are generally sequential, from left to right, but can also occur in parallel, and be iterative.

These sub-processes are:

7.1. Manage data and metadata repositories - This sub-process focuses on managing the systems where data and metadata are stored for dissemination purposes, including:

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

7.2. Produce products - This sub-process produces the products, as previously designed, to meet user needs. The products can take many forms including printed publications, press releases and internet databases. Typical steps include

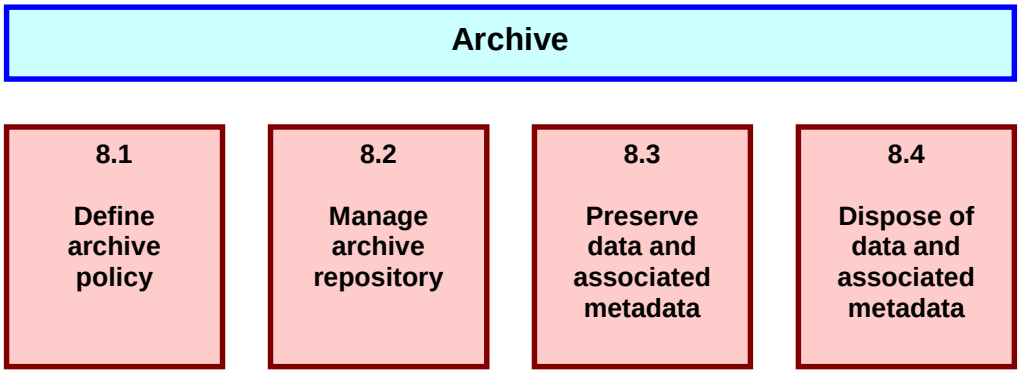
- setting up the product templates and other required product components;
- preparing the product components;
- adding the content to the product;
- editing and checking the product meets publication standards.

7.3. Manage release of products - 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.

7.4. Market and promote products - This sub-process concerns the active promotion and marketing of statistical products to help them reach the widest possible audience. It includes the use of customer relationship management tools (which may be similar to the provider relationship management tools in sub-process 4.2), to help to build market intelligence, and to ensure that the “brand” of official statistics produced by the agency is familiar to users. It also includes the establishment and use of tools and systems (e.g. web sites, wikis, blogs, etc.) to facilitate the process of communicating statistical information to users.

7.5. Manage customer queries - This sub-process ensures that customer queries are recorded, and that responses are provided within agreed deadlines. These queries should be regularly reviewed to provide an input to the over-arching quality management process, as they can indicate new or changing user needs.

Phase 8 – Archive



This phase manages the archiving and disposal of statistical data and metadata. Given the reduced costs of data storage, it is possible that the archiving strategy adopted by a statistical agency does not include provision for disposal, so this sub-process may not be relevant for all agencies. For statistical outputs produced regularly, archiving occurs in each iteration, however defining the archiving policy is likely to occur less regularly.

This phase is made up of four sub-processes, which are generally sequential, from left to right, but can also occur in parallel, and be iterative. These sub-processes are:

8.1. Define archive policy – This sub-process is where the archiving policy for the statistical data and metadata resulting from a statistical business process is determined. This policy should also consider the archiving of intermediate outputs such as the sample file, the raw data from the collect phase, and the results of the various stages of the process and analyse phases. The archive policy for a specific statistical business process may be fully or partly dependent on the more general archiving policy of the statistical agency, or, for national agencies, the government sector. The policy should include consideration of the medium and location of the archive, as well as the requirement for keeping duplicate copies. It should also consider the conditions (if any) under which data and metadata should be disposed of.

8.2. Manage archive repository – This sub-process concerns the management of one or more archive repositories. These may be databases, or may be physical locations where copies of data or metadata are stored. It includes:

- maintaining catalogues of data and metadata archives, with sufficient information to ensure that individual data or metadata sets can be easily retrieved;
- testing retrieval processes;
- periodic checking of the integrity of archived data and metadata.

8.3. Preserve data and associated metadata – This sub-process is where the data and metadata from a specific statistical business process are archived. It includes:

- identifying data and metadata for archiving in line with the archive policy defined in 8.1;
- formatting those data and metadata for the repository;
- loading or transferring data and metadata to the repository;
- cataloguing the archived data and metadata;
- verifying that the data and metadata have been successfully archived.

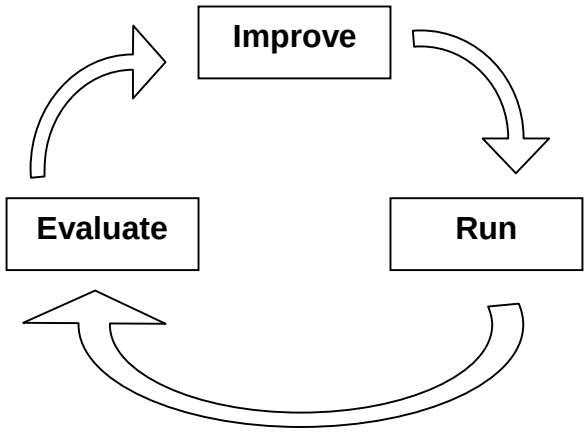
8.4. Dispose of data and associated metadata – This sub-process is where the data and metadata from a specific statistical business process are disposed of. it includes;

- identifying data and metadata for disposal, in line with the archiving policy defined in 8.1;
- disposal of those data and metadata;
- recording that those data and metadata have been disposed of.

Over-arching processes

Quality Management

This process is present throughout the model, in that each phase and sub-process should be evaluated, ideally at every iteration, but at least according to an agreed schedule. These evaluations should result in feedback, which should be used to improve the relevant sub-process(es), creating a quality loop:



Quality management can take several forms, including:

- Seeking and analysing user feedback;
- Reviewing operations and documenting lessons learned;
- Benchmarking / peer reviewing processes with other organizations.

Evaluation will normally take place within an institutional quality framework, and may therefore take slightly different forms and deliver slightly different results from agency to agency. This stresses the importance of the benchmarking and peer review approaches to evaluation, and whilst these approaches are not feasible for every iteration of every statistical business process, they should be used in a systematic way according to a pre-determined schedule.

Metadata Management

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. The key challenge is to ensure that they are captured, stored and transferred from phase to phase alongside the data they refer to. A metadata management strategy and system(s) are therefore vital to the operation of this model.

Part A of the Common Metadata Framework identifies the following twenty core principles for metadata management, all of which are intended to be covered in the over-arching Metadata Management process:

1. Make metadata-related work an integral part of business processes across the organization.

2. Describe metadata flow with the statistical and business processes (alongside the data flow and business logic).

3. Ensure that customers are clearly identified for all metadata processes, and that all metadata capturing will create value for stakeholders.

4. Ensure that metadata presented to the end-users match the metadata that drove the business process or were created during the business process.

5. Develop a statistical metadata system (SMS) as a self-sustainable project, independent of any e-production systems.

6. Ensure the SMS is the definitive set of tools, stores and services to support metadata use and further development in a statistical agency. If a metadata store, tool or service is not defined by the management to be a part of the SMS, then it is not an “approved” metadata facility.

7. Recognise the diversity of metadata, with different views corresponding to the different uses to which the data are being put. Different users require different levels of detail. Metadata appear in different formats depending on the processes and goals for which they are produced and used.

8. Make metadata active to the greatest extent possible. Active metadata drive other processes and actions will therefore be accurate and up-to-date.

9. Manage metadata with a life-cycle focus (including maintenance and update).

10. Preserve history (old versions) of metadata.

11. Capture metadata at their natural sources, preferably automatically as bi-product of other processes. Minimize errors by entering only once where possible.

12. 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 choice of standard exchange language, such XML.

13. Ensure that all data and other objects of the SMS are well supported by accessible metadata that are of appropriate quality.

14. Ensure that metadata are readily available and useable in the context of client's information needs (whether client is internal or external).

15. Ensure that there is a single, authoritative source ('registration authority') for each metadata element.

16. Associate a registration process (workflow) with each metadata element, so that there is a clear identification of ownership, approval status, date of operation etc.

17. Reuse metadata where possible for statistical integration as well as efficiency reasons (no new metadata elements are created until the designer/architect has determined that no appropriate element exists and this fact has been agreed by the relevant 'standards area').

18. Implement a cost/benefit mechanism to ensure that the cost to producers of metadata is justified by the benefit to users of metadata.

19. Ensure that variations from standards are tightly managed/approved, documented and visible.

20. Ensure systematic training and transfer of know how for all partners involved. Train trainers.