

Distr.
GENERAL

Working Paper No.10
28 March 2008

ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE
EUROPEAN COMMUNITIES (EUROSTAT)

ORGANISATION FOR ECONOMIC
COOPERATION AND DEVELOPMENT
(OECD)
STATISTICS DIRECTORATE

Joint UNECE/Eurostat/OECD meeting on management of statistical information systems (MSIS)
(Luxembourg, 7-9 April 2008)

PROPOSAL FOR A NEW GENERIC STATISTICAL BUSINESS PROCESS MODEL

Supporting Paper

Prepared by the UNECE Secretariat ¹

I. BACKGROUND

1. Part C of the [Common Metadata Framework \(CMF\)](#) – Metadata and the Statistical Cycle – refers to the phases of the statistical cycle (or “statistical business process”) and provides generic terms to describe them. The intention is for statistical organizations to agree on standard terminology to aid their discussions on developing statistical metadata systems. In 2006, when the structure for Part C was first being developed, these phases were proposed to be:

- (1) survey planning and design;
- (2) survey preparation;
- (3) data collection;
- (4) input processing;
- (5) derivation, estimation, aggregation;
- (6) analysis;
- (7) dissemination;
- (8) post survey evaluation.

2. During the [workshop on Part C of the CMF](#), held in July 2007, five national statistical offices² reported on the terms currently used to describe phases of the statistical business process within their organization. This led to a discussion about the generic terms proposed for the CMF, mentioned above. The meeting participants agreed that with the addition of ‘Archive’ and ‘Evaluate’ phases, the model currently used by Statistics New Zealand would provide a better basis for the generic CMF model. This model is 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. Using the work of Statistics New Zealand as a basis, the METIS Steering Group has developed a proposal for a Generic Statistical Business Process Model (see Annex 1 for a graphical representation and Annex 2 for a more detailed description). The model can be divided into three tiers: the highest level simply names the nine phases of the statistical business process (also known as the statistical life-cycle or value chain); the second level identifies the sub-processes within each phase; and the third level identifies the sub-processes within each second level component. The model is intended to apply to statistical production regardless of the

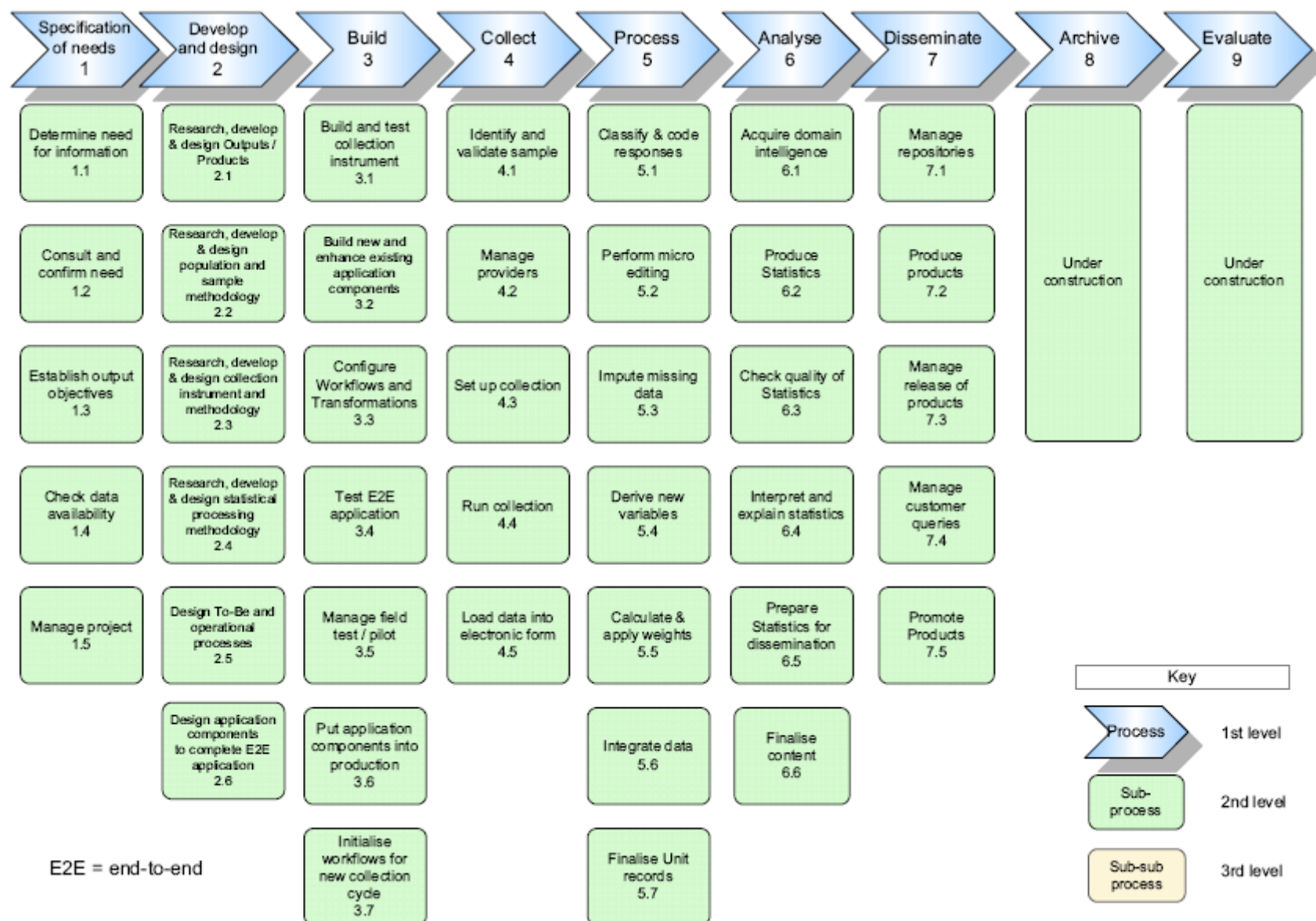
¹ Prepared by Jessica Gardner (jessica.gardner@unece.org) and Steven Vale (steven.vale@unece.org), based largely on material kindly provided by Statistics New Zealand

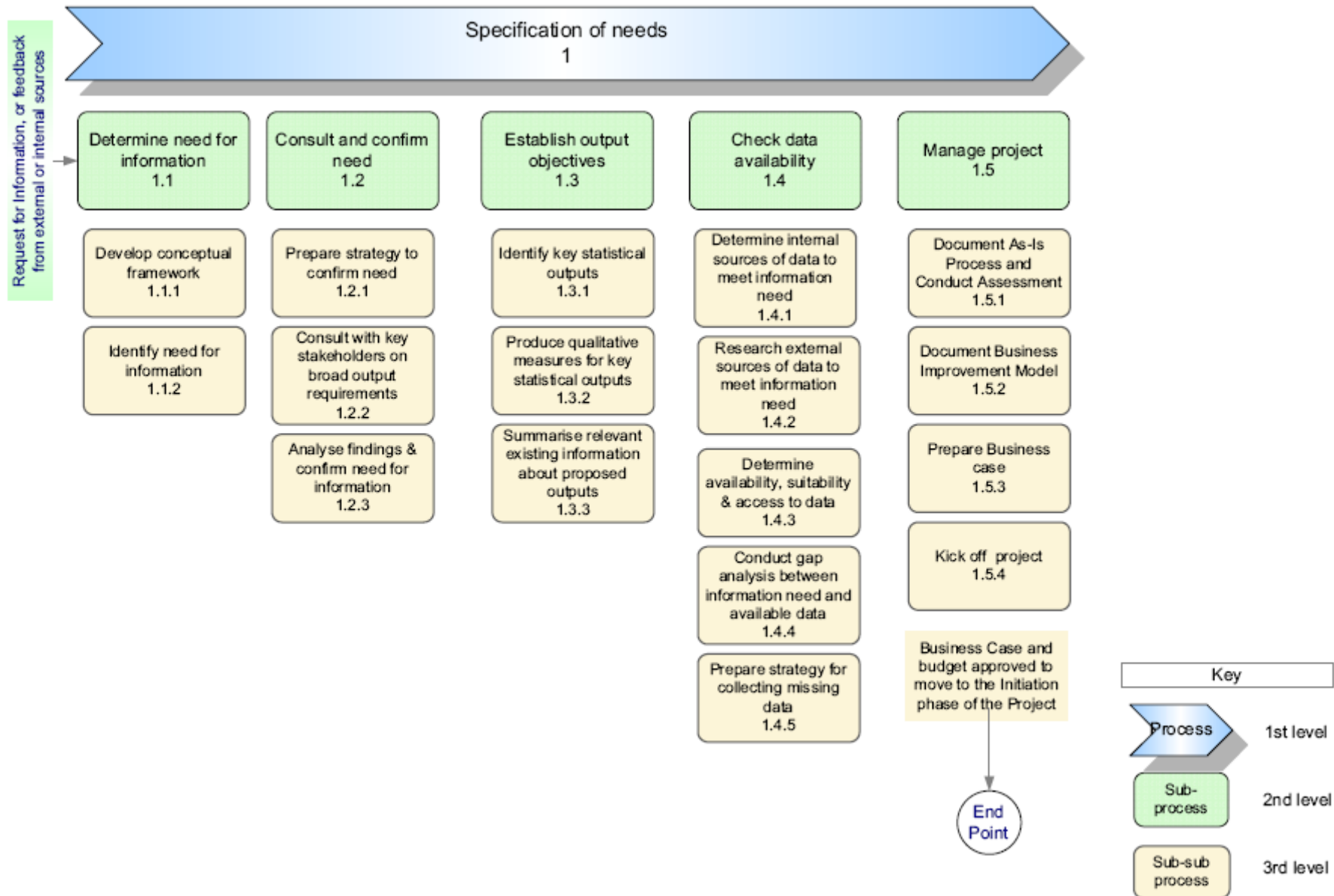
² This was a component of the case studies presented by Canada, Ireland, New Zealand, South Africa and Sweden.

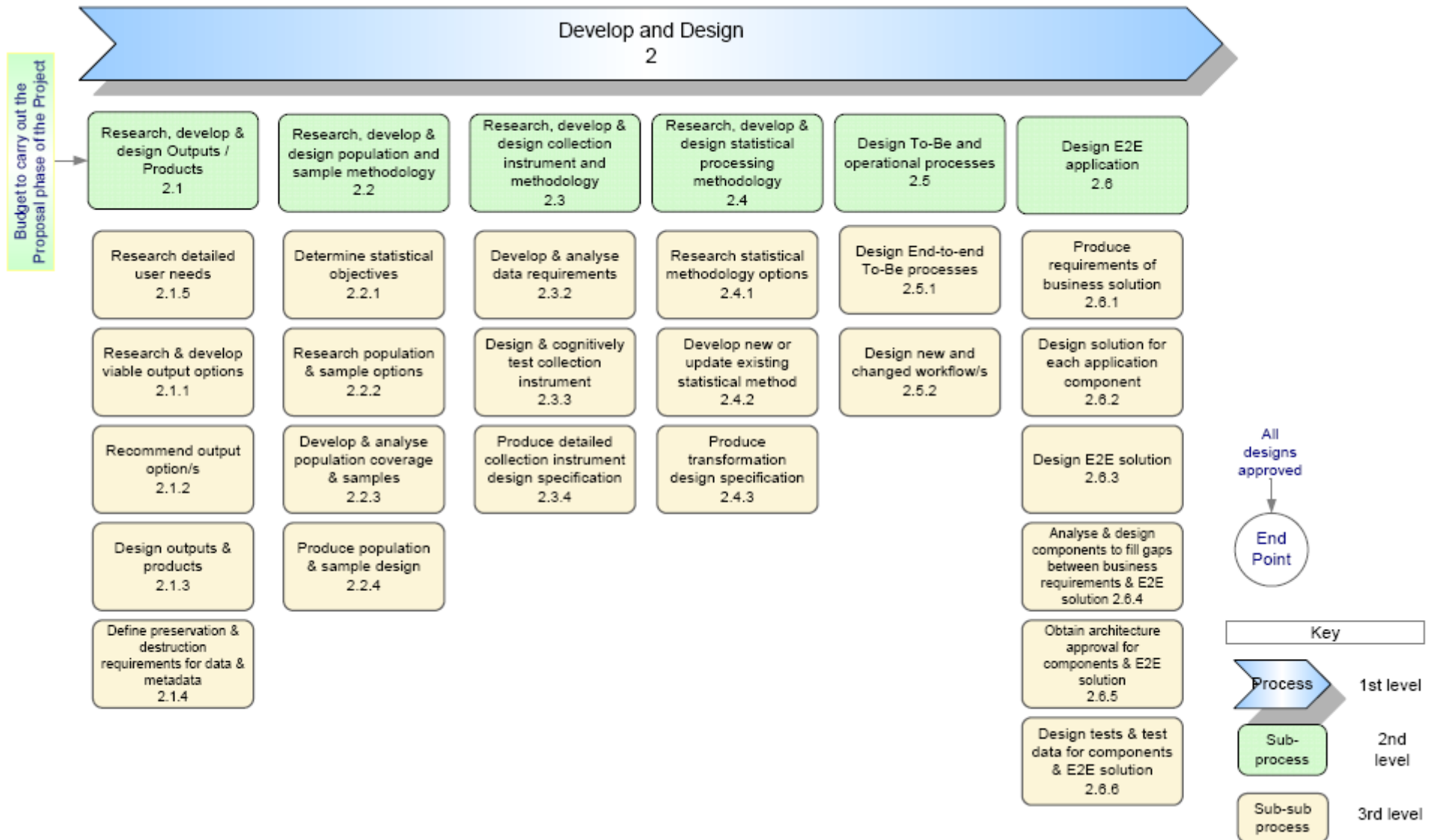
data source (surveys, administrative records, data integration etc.), it also encompasses data quality and the production of metadata.

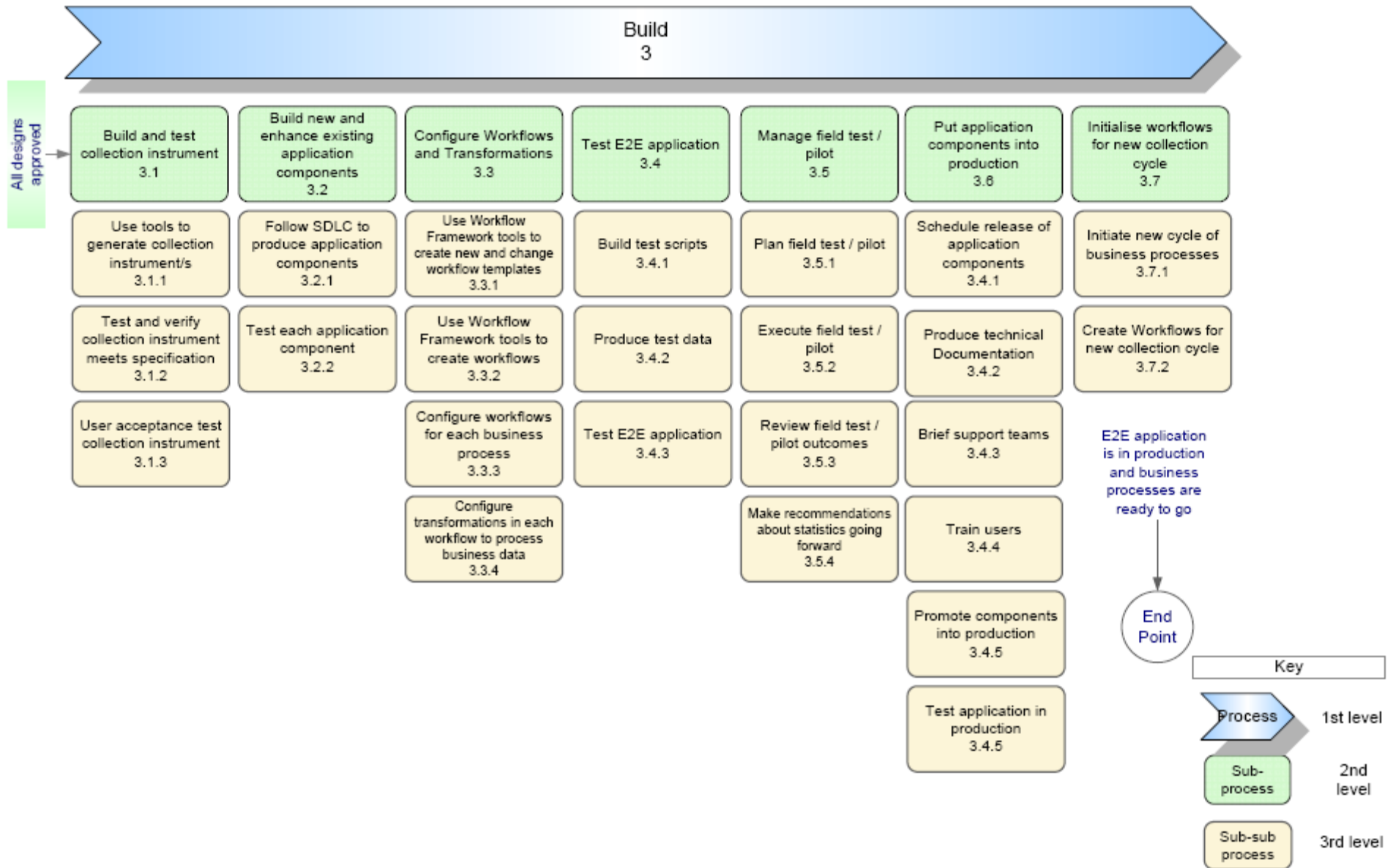
4. Further refinements are likely to be needed to make the model more generic, and it is acknowledged that some statistical business processes will not use all phases of the model. If the METIS Work Session approves the principle of having such a model, further drafts will be produced and circulated, taking account of specific comments received.

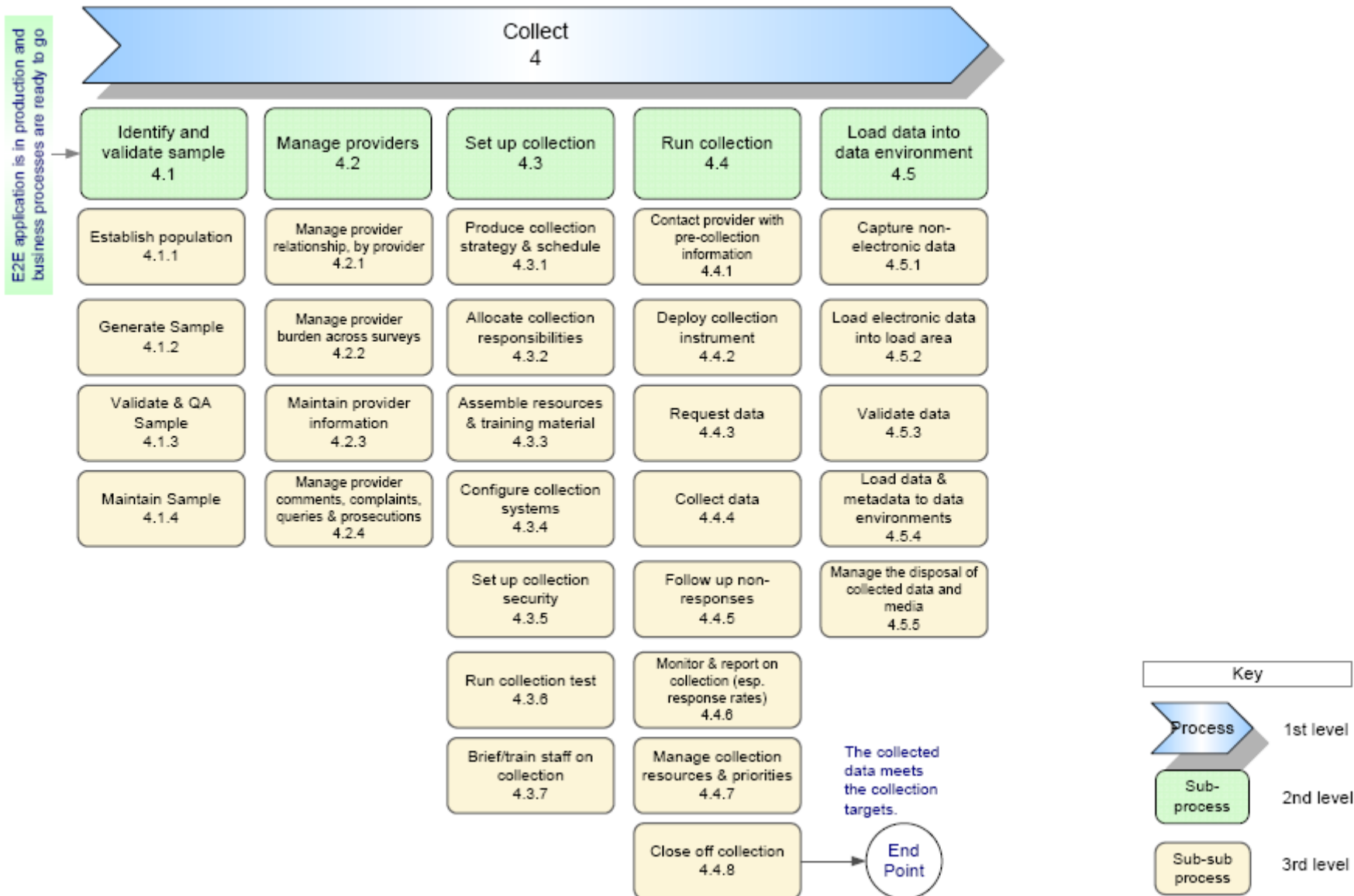
Annex 1 - The Generic Statistical Business Process Model - DRAFT -

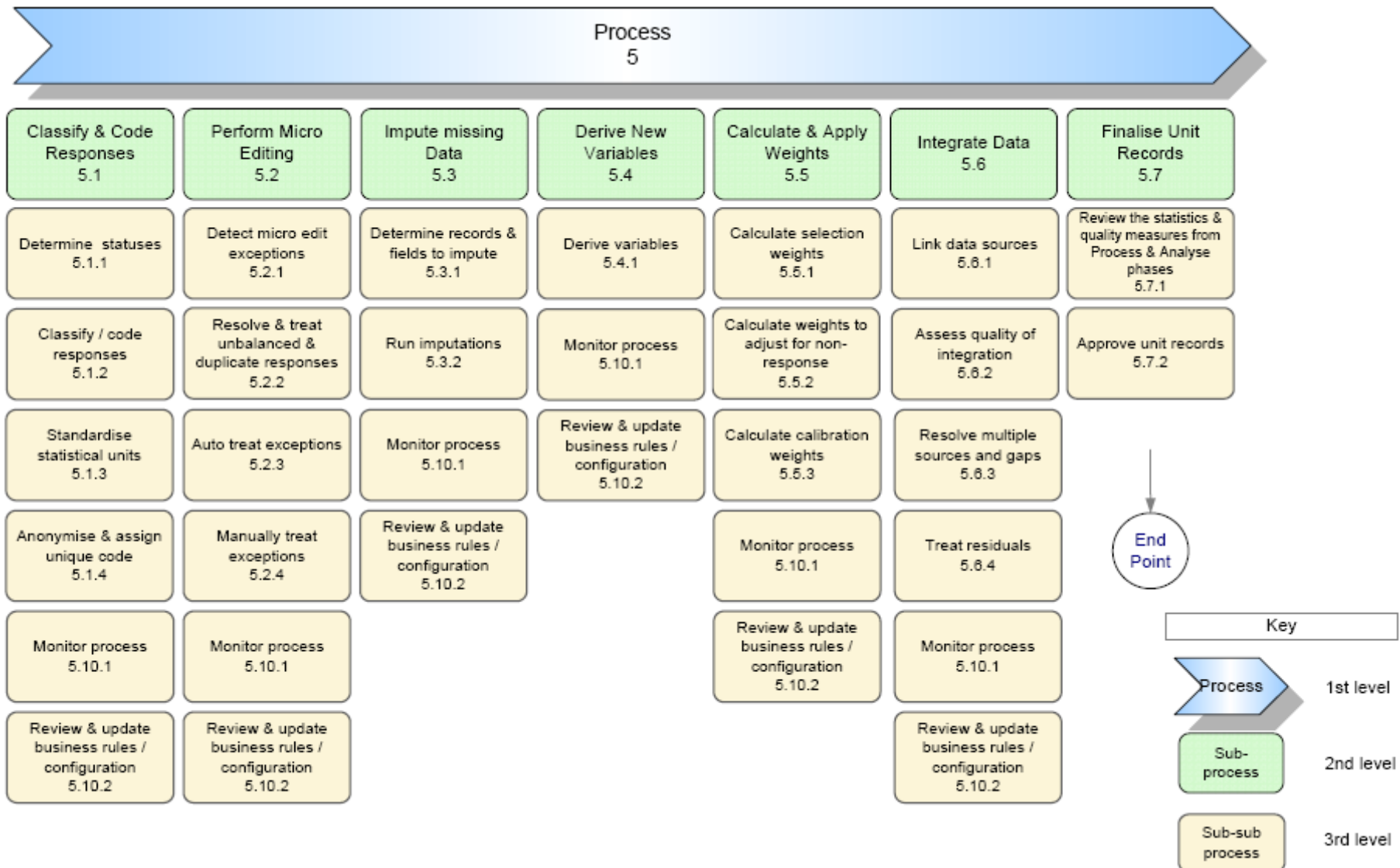


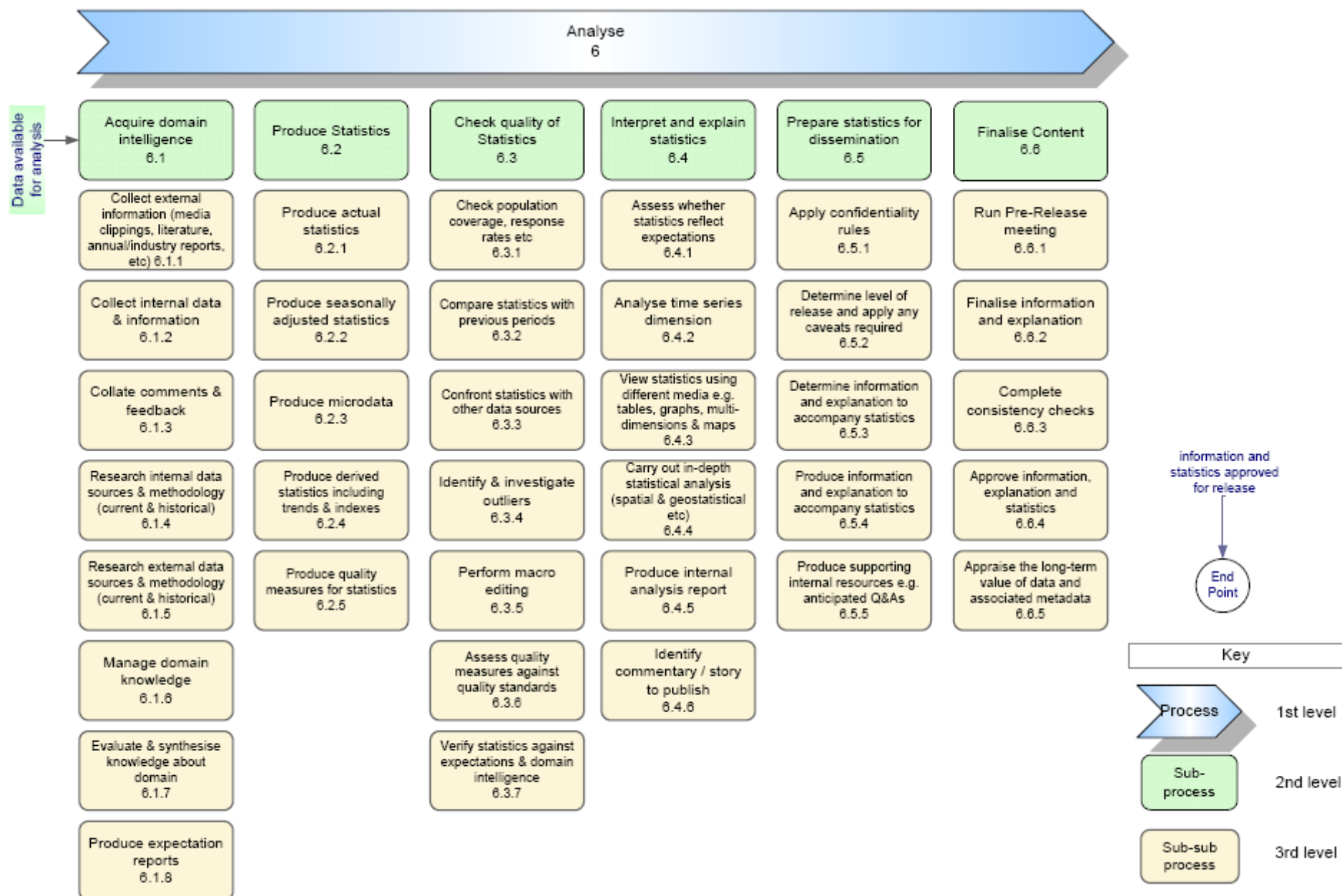


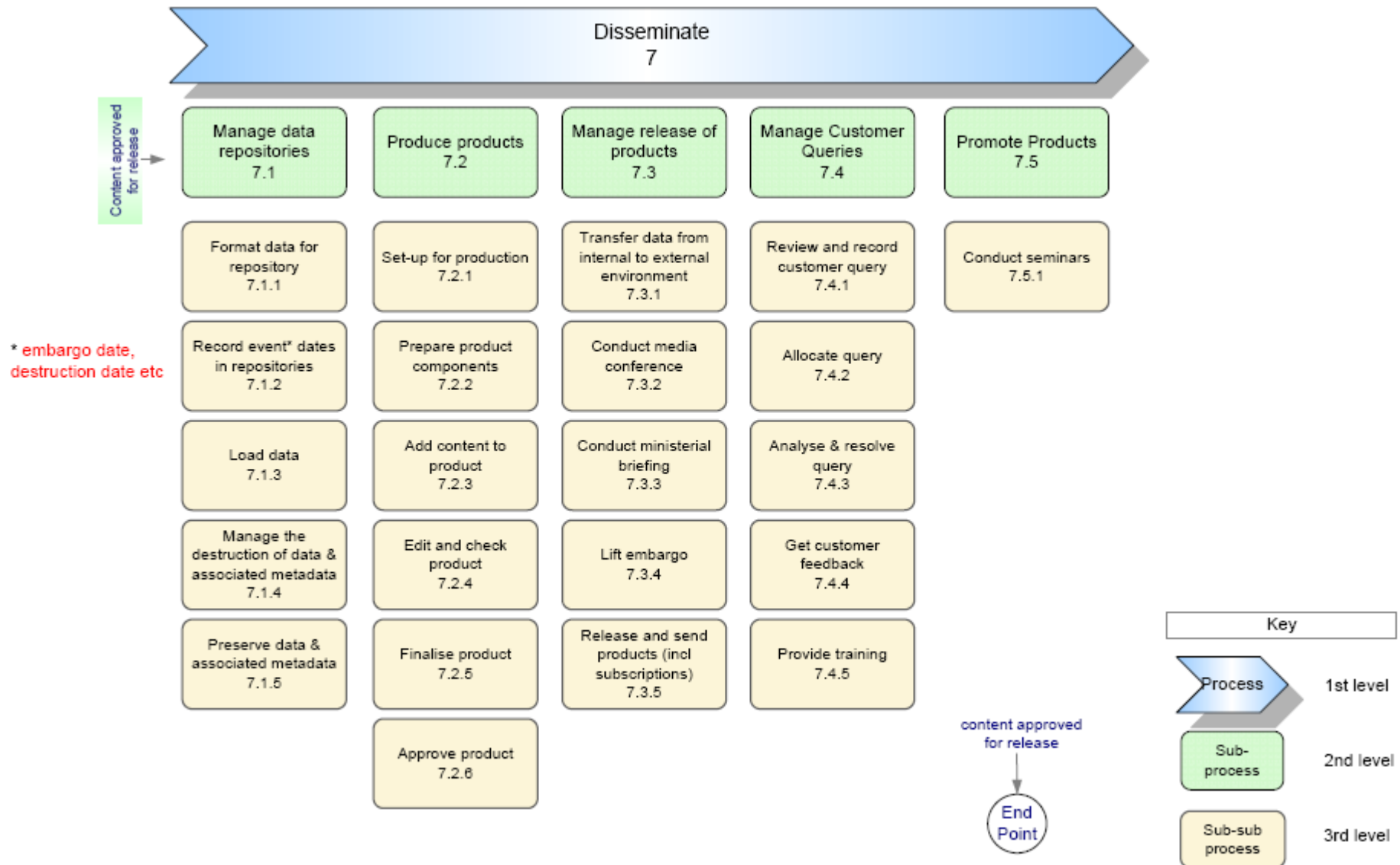












Annex 2 - The Generic Statistical Business Process Model - Description

Phase 1: Specification of 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.

For statistical outputs produced on a regular basis, this phase occurs for the first iteration, and any subsequent reviews.

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. International Statistics are looked at, along with the methodology other Statistical Offices use to produce and disseminate these.

1.2 Consult and confirm need - This sub-process focuses on consulting with the stakeholders and confirming the need for the statistics. A good understanding of this is needed, so that the statistical agency knows what it is expected to deliver. Understanding what the stakeholders want (or making sure that existing needs are still current) is the critical part of this sub-process.

1.3 Establish output objectives - This sub-process focuses on understanding the statistics required to meet the needs of the stakeholders. Research and analyses are done on the statistical requirements and their quality measures through consultation with the stakeholders etc.

1.4 Check data availability - This sub-process focuses on identifying one or more sources of data (and their availability) proposed to deliver the statistics. Once existing data sources and availability have been assessed, a strategy for filling the gaps is prepared.

1.5 Manage project - This sub-process focuses on determining how a project will manage the statistical production process. To do this, the sub-process includes:

- Business process modelling activities to document the “As-Is” business process (if it already exists), with details on how the current statistics are produced. The document highlights inefficiencies and issues to be addressed, as well as opportunities for moving forward.
- Business improvement model, with the “To-Be” solution, detailing the people, processes and technology required to produce the new or reviewed statistics.

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 occurs for the first iteration, and any subsequent reviews.

This 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. These sub-processes are:

2.1 Research, develop and design outputs/products - The starting point for this research and development work is the metadata from similar or previous collections. These metadata provide the framework for defining the data elements required to produce the statistical outputs and products. The input data elements are sourced

from either survey questions or existing data sources. These data sources can be a combination of one or more internal collections or external administrative data.

2.2 Research, develop and design population and sample methodology - This sub-process focuses on researching, developing and designing the population and sample methodology, or for data integration projects, determining the population coverage. The actual sample is created in Collect (sub-process 4.1: Identify and validate sample), using the methodology, specified in this sub-process.

2.3 Research, develop and design collection instrument and methodology - The actual activities in this sub-process vary according to the type of collection instruments required. The collections instruments 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 enables the reuse of questions, and related attributes, and a questionnaire tool, which enables the quick and easy compilation of questions into formats available for early 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 Research, develop and design 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).

2.5 Design To-Be and operational processes - This sub-process is where the business processes for the collection are defined.

2.6 Design E2E application

This sub-process initially, completes the gap analysis between the detailed business processes and the existing end-to end (E2E) components, to determine what items are missing and need to be built. Existing components are examined to ensure they are fit for purpose for the collection in question. A general principle is to reuse processes and technology across the statistical value chain.

Phase 3: Build

This phase includes:

- building and testing the collection instruments
- building, enhancing and testing the components designed for the end-to-end solution, and
- putting the components into production.

For statistical outputs produced on a regular basis, this phase occurs for the first iteration, and any subsequent reviews.

The “Build” phase is broken down into seven 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 Build and test collection instruments - 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 uses one or more collection modes to receive the data, e.g. interviewers to complete the questions, either in person or over the telephone or providers complete a survey, either on paper or on the web. Collection instruments may also be data extraction routines used to gather data from existing statistical or administrative data sets.

3.2 Build new and enhance existing application components - This sub-process describes the activities to build new and enhance existing software components needed for the E2E solution, as designed in the “Develop and design” phase. Components may include dashboard functions and features, data repositories, transformations, workflow framework components and metadata.

3.3 Configure workflows and transformations - This sub-process configures the workflows and transformations, used to systemise the business processes, from receiving the data, right through to evaluating the final statistical outputs.

3.4 Test E2E application - This sub-process tests the E2E (end-to-end) solution, as though it is in production, and ensures it supports the business process. The E2E solution is a combination of people (roles), processes and technology components.

3.5 Manage field test / pilot - This sub-process is optional, it describes the activities to manage a field test or pilot where it is required, before the new E2E solution is used live. Components tested can include new technology, the mode(s) of collection, collection procedures, data integration and other processing methodologies.

3.6 Put application components into production - This sub-process includes the activities to put the newly built technology components into production ready for use by business areas. The activities include:

- producing documentation about the technology components, including technical documents
- producing documentation about the E2E solution i.e. user manuals
- training the business users on how to use the E2E solution
- moving the technology components into production, and
- ensuring the E2E solution works in the production environment.

3.7 Initialise workflows for new collection cycle - This sub-process describes the build activities carried out to initialise the workflows for the new cycle, and to configure the workflows and transformations to make them unique to this collection cycle.

Phase 4: Collect

This phase collects all external data, using different collection modes, and loads it into the appropriate data environment. Note: 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 Identify and validate sample - This sub-process creates the sample for this iteration of the collection. Once the sample is approved (a manual step), the “Collect” phase begins in earnest.

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:

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

4.3 Set up collections - 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 set up.

4.4 Run collection - This sub-process is where the collection is implemented, with the different collection instruments used to collect the data. Note: 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 the collection is closed and a report on the collection is produced.

4.5 Load data into electronic form - This sub-process validates and loads the data (and metadata) into the a suitable electronic form. Some data, e.g. unstructured data, may require more checks and manual intervention than other data.

Phase 5: Process

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 “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 an early picture of the statistics to be built up, and increases the time available for analysis.

This phase is broken down into seven 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 respondents - This sub-process classifies and codes the input data, and is 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 Perform micro-editing - The Micro-editing sub-process applies to each record, and looks at the unit record data in the context to that record only. It is usually run iteratively, and is executed by one or more workflows, to validate data against predefined edit rules, apply auto-edits, raise alerts and manually edit the data. Micro-editing applies to unit records from all types of collections, before and after data integration.

5.3 Impute missing data - This sub-process is usually run iteratively, and describes the activities, executed by workflows, to identify missing data, both at the variable level (item non-response), and at the unit level (unit non-response). It includes:

- the manual selection of data to include or remove from the imputation formula
- the imputation of data using the configured method e.g. “hot-deck”, “cold-deck” or other methods
- 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 runs the workflow that derives the predefined derived variables. It is used to create variables that are not explicitly provided in the collection and are needed to deliver the required outputs.

5.5 Calculate and apply weights - This sub process creates and applies the weights to unit data records according to the methodology created in sub-process 2.4: Research, develop and design statistical processing methodology.

5.6 Integrate data - This sub-process integrates one or more sources of data. The input data can be from a mixture of external or internal data sources, and a variety of collection modes.

5.7 Finalise unit records - This sub-process is responsible for verifying that the clean unit records are “fit for purpose” (i.e. can deliver the statistics required by the stakeholders), and finally approving them.

Phase 6: Analyze

In this phase, statistics are produced, examined in detail, interpreted, understood and made ready for dissemination. This phase describes 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 sub-processes and activities are generic for all statistical outputs, regardless of how the data were sourced.

The Analyse process 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-processes includes many ongoing activities involved with the gathering of intelligence, with the cumulative effect of building up the body of knowledge about the 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. This means that when it comes to sub-process 6.4: Interpret and explain statistics a better explanation can be given about the message the statistics convey.

6.2 Produce statistics - This sub-process is where domain intelligence is applied to the data collected to produce statistical outputs.

6.3 Check quality of 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 an in-depth statistical analyses.

6.5 Prepare statistics for dissemination - This sub-process prepares the statistics and associated information for dissemination by:

- applying the confidentiality rules
- determining the level of release, and applying caveats
- producing the supporting information to accompany the statistics, and
- 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 up to the required quality level. The activities include:

- running an optional pre-release meeting with related internal subject matter experts
- finalising the information and explanation
- completing consistency checks
- approving the statistical content
- appraising the long term value of all the data and associated metadata.

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 repositories - This sub-process focuses on managing the data repositories, including:

- formatting the data ready to be put into the data repositories
- loading the data (and relevant metadata) into the repositories
- preserving the data and required metadata.

7.2 Produce products - This sub-process produces the products, previously designed, to fulfil the stakeholders' 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, and
- 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.

7.4 Manage customer queries - This sub-process ensures that customer queries are reviewed and recorded, and that responses are provided.

7.5 Promote products - This sub-process markets statistical products to help them reach the widest possible audience.

Phase 8: Archive

This phase is still being defined, but is likely to include the following sub-processes:

- Ensure long-term storage of data and metadata
- Ensure back-up copies are taken and stored separately
- Periodically test that data and metadata can be successfully retrieved from archive and back-up copies
- Ensure archived data are catalogued

Phase 9: Evaluate

This phase is still being defined, but is likely to include the following sub-processes:

- Seek, analyze and act on user feedback
- Review one-off operations and document lessons learned
- Carry out scheduled reviews of regularly repeated operations
- Benchmark / peer review operations with other organizations