Exploring the relationship between DDI, SDMX and the Generic Statistical Business Process Model

Steven Vale, United Nations Economic Commission for Europe
steven.vale@unece.org

0. Abstract

The UNECE and the Conference of European Statisticians Steering Group on Statistical Metadata (better known as "METIS") have recently developed the "Generic Statistical Business Process Model" (GSBPM). This model has already been widely adopted by national statistical organisations around the world, and is intended to facilitate the convergence of statistical production processes, both within and between organisations. There are certain obvious similarities between the GSBPM and the DDI 3 Combined Life Cycle Model.

At the same time, there is growing interest in official statistics in using DDI 3 in the earlier phases of the statistical production process (particularly for microdata), perhaps in combination with SDMX (Statistical Data and Metadata eXchange) standards, which are seen as more appropriate for macrodata. This paper highlights the work so far on exploring the relationships and interoperability between DDI, SDMX and the GSBPM, as a way of modernising and standardising (i.e., “industrialising”) statistical production. It was presented at the 2nd Annual European DDI Users Group Meeting in Utrecht, Netherlands, in December 2010.

1. Introducing the GSBPM

The original aim of the GSBPM was to provide a basis for statistical organisations to agree on standard terminology to aid their discussions on developing statistical metadata systems. It was conceived as a flexible tool to describe and define the set of business processes needed to produce official statistics. The GSBPM is, however, increasingly being used in other contexts such as harmonising statistical computing infrastructures, facilitating the sharing of software components, and providing a framework for process quality assessment and improvement.

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

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

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

The GSBPM is not intended to be a rigid framework in which all steps must be followed in a strict order, but rather a model that identifies the steps in the statistical business process, and the interdependencies between them. It aims to be sufficiently generic to be widely applicable, and to encourage a standard view of the statistical business process, without becoming either too restrictive or too abstract and theoretical. Different business processes will follow different paths through the model, using different processes, sometimes in a different order.

The GSBPM comprises four levels:

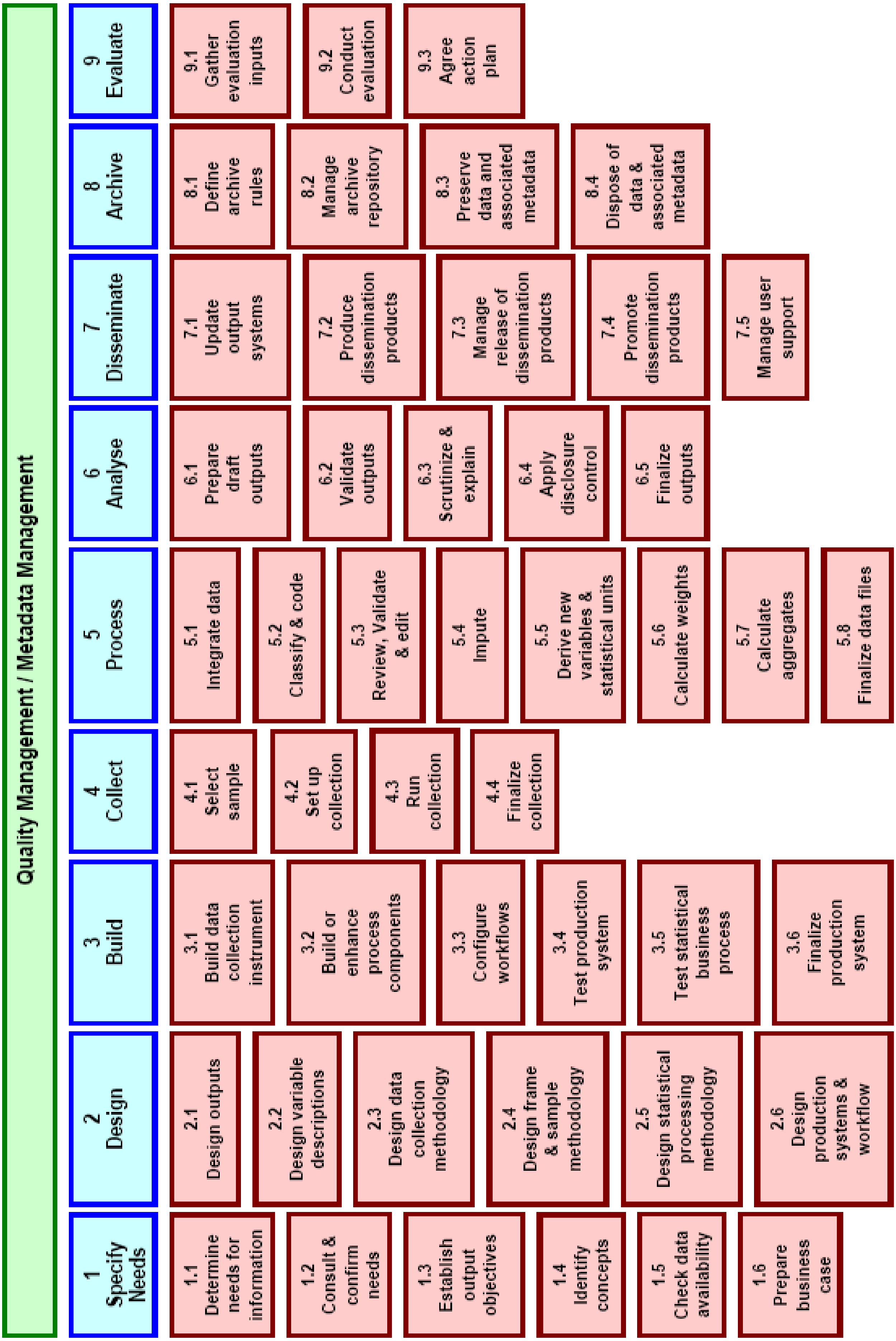
- Level 0, the statistical business process
- Level 1, the nine phases of the statistical business process
- Level 2, the sub-processes within each phase
- Level 3, a description of those sub-processes

Levels 1 and 2 are illustrated in Figure 1. Information about level 3 can be found in the GSBPM documentation. Further levels of detail may be appropriate for certain statistical business processes or in certain organisations, but these are unlikely to be sufficiently generic to be included in this model.

The GSBPM also recognises several overarching processes that apply throughout the nine phases, and across statistical business processes. These can be grouped into two categories, those that have a statistical component, and those that are more general, and could apply to any sort of organisation. Examples of overarching statistical processes include:

- Quality management – including quality assessment and control mechanisms
- Metadata management – ensuring that metadata retain their links with data throughout the business process
- Statistical framework management – developing standards, methodologies, concepts and classifications that apply across multiple processes
- Human resource management
- Financial management
- Project management

Figure 1: GSBPM phases and sub-processes



Quality Management / Metadata Management

1 Specify Needs
- 1.1 Determine needs for information
- 1.2 Consult & confirm needs
- 1.3 Establish output objectives
- 1.4 Identify concepts
- 1.5 Check data availability
- 1.6 Prepare business case

2 Design
- 2.1 Design outputs
- 2.2 Design variable descriptions
- 2.3 Design data collection methodology
- 2.4 Design frame & sample methodology
- 2.5 Design statistical processing methodology
- 2.6 Design production systems & workflow

3 Build
- 3.1 Build data collection instrument
- 3.2 Build or enhance process components
- 3.3 Configure workflows
- 3.4 Test production system
- 3.5 Test statistical business process
- 3.6 Finalize production system

4 Collect
- 4.1 Select sample
- 4.2 Set up collection
- 4.3 Run collection
- 4.4 Finalize collection

5 Process
- 5.1 Integrate data
- 5.2 Classify & code
- 5.3 Review, Validate & edit
- 5.4 Impute
- 5.5 Derive new variables & statistical units
- 5.6 Calculate weights
- 5.7 Calculate aggregates
- 5.8 Finalize data files

6 Analyse
- 6.1 Prepare draft outputs
- 6.2 Validate outputs
- 6.3 Scrutinize & explain
- 6.4 Apply disclosure control
- 6.5 Finalize outputs

7 Disseminate
- 7.1 Update output systems
- 7.2 Produce dissemination products
- 7.3 Manage release of dissemination products
- 7.4 Promote dissemination products
- 7.5 Manage user support

8 Archive
- 8.1 Define archive rules
- 8.2 Manage archive repository
- 8.3 Preserve data and associated metadata
- 8.4 Dispose of data & associated metadata

9 Evaluate
- 9.1 Gather evaluation inputs
- 9.2 Conduct evaluation
- 9.3 Agree action plan

2. Links to DDI and SDMX

SDMX

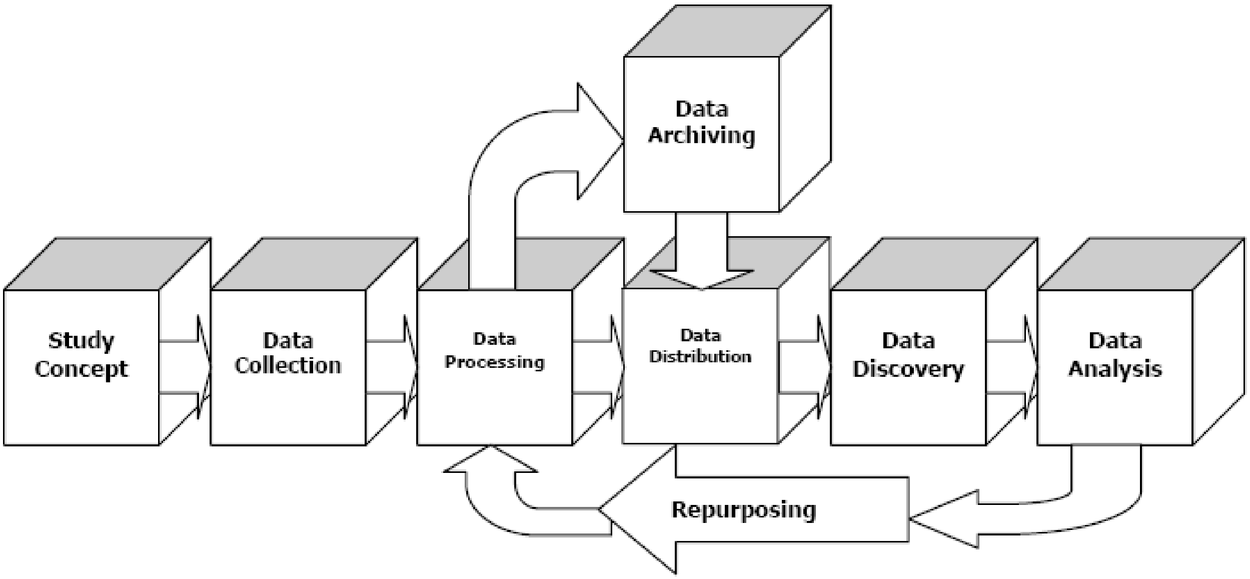
The SDMX (Statistical Data and Metadata eXchange) standards do not provide a model for statistical business processes in the same sense as the GSBPM. However they do provide standard terminology for statistical data and metadata, as well as technical standards and content-oriented guidelines for data and metadata transfer, which could be applied between sub-processes within a statistical organisation.

The relationship between the GSBPM and SDMX was discussed at the April 2008 meeting of the METIS group. The final report of that meeting records a suggestion to incorporate the model into the SDMX content-oriented guidelines as a cross-domain concept. It was also discussed at the 2009 SDMX Global Conference, where the possibility was raised that SDMX could provide the format for data transmission between sub-processes, within a statistical organisation[1]. These issues will be considered further by the SDMX Statistical Working Group, which will start its work in 2011.

The DDI 3 Combined Life Cycle Model

This model (see Figure 2) 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 is rather different to the GSBPM, which specifically applies to official statistical organisations. 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.

Figure 2: The DDI 3 Combined Life Cycle Model



Source: Data Documentation Initiative (DDI) Technical Specification, Part I: Overview, Version 3.1, October 2009, http://www.ddialliance.org.

[1] See: Vale S. and Hamilton A. “The Statistical Business Process View: A Useful Addition to SDMX?” (2009), available at: http://www1.unece.org/stat/platform/display/metis/Papers+about+the+GSBPM

The main differences between the GSBPM and the DDI Combined Life Cycle Model are:

- The GSBPM generally assumes that the whole process is carried out by one organisation (though for large processes such as censuses, certain sub-processes may be sub-contracted). The DDI model seems to recognise that steps such as “Data analysis” and “Repurposing” may be carried out by different organisations to the one that collected the data. This reflects a fundamental difference between practices in the research and official statistics communities, with the research community having more scope for collaboration between organisations during the production process.
- The DDI model replaces the dissemination phase with “Data Distribution” which takes place before the analysis phase. This reflects the 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 harmonised data set. This generally refers to some re-use of a data-set that was not originally foreseen in the design and collect phases. In the GSBPM, if outputs from one process are re-used for another purpose, these are treated as two separate processes (two instances of the model). The second process identifies the data in phase 1 (Specify Needs) where there is a sub-process to check the availability of existing data, and obtains them in phase 4 (Collect) and then uses them to produce new outputs.
- The DDI model has separate phases for data discovery and data analysis, whereas these functions are combined within phase 6 (Analysis) in the GSBPM. In some cases, elements of the GSBPM analysis phase may also be covered in the DDI “Data Processing” phase, depending on the extent of analytical work prior to the “Data Distribution” phase.
- The GSBPM places data archiving towards the end of the process, after the analysis phase (though as the GSBPM is not a linear model, archiving can in practice take place at multiple points in the process). This is probably more of a presentational issue than a major conceptual difference.
- The GSBPM explicitly identifies “overarching processes”, such as quality and metadata management, whereas these are more implicit in the DDI model.

It is clear the GSBPM and the DDI Combined Life Cycle Model serve slightly different purposes. This provides a good justification for the differences between them. However, despite these differences in purpose, there are also a lot of similarities. It is therefore useful to map the two models to try to get a better understanding of how they might interact. The GSBPM documentation includes an initial, high-level mapping (shown in Figure 3), but further work on a more detailed mapping is needed.

Figure 3: Mapping the Models

Generic Statistical Business Process Model	DDI 3.0 Combined Life Cycle Model
1 Specify Needs	Study Concept Repurposing (part)
2 Design	
3 Build	
4 Collect	Data Collection
5 Process	Data Processing (mostly) Repurposing (part)
6 Analyse	Data Discovery Data Analysis Data Processing (part)
7 Disseminate	Data Distribution
8 Archive	Data Archiving
9 Evaluate	
Quality Management	
Metadata Management	

3. Combining Standards to Industrialise Official Statistics

The GSBPM is increasingly being used by national and international statistical organisations as a framework to help to develop and modernise the production of official statistics. This is part of a general trend towards a more process-oriented approach to producing statistics, sometimes referred to as the "industrialisation of statistical production". The basic idea is that statistical organisations can not continue to afford developing methods, processes and tools separately for different statistical domains. Cost and efficiency pressures are leading inevitably towards an "industrial revolution" bringing in standard systems and approaches for statistical production regardless of the topic.

The GSBPM helps to identify the different sub-processes to be harmonised, and to encourage a modular approach to the development of statistical methods and software tools. For example, can it really be justified to have ten different tools for imputing missing data, when the process is conceptually the same? A related UNECE initiative is using the GSBPM as a classification system for an inventory of statistical software tools that can be shared between organisations[2]. This reinforces the modular approach to statistical production by providing further cost incentives resulting from sharing development work between organisations.

However, the GSBPM is not sufficient in itself to implement this sort of change. Other standards are needed, for example, an information model, and data and metadata format standards. These are required to ensure efficient flows between sub-processes, and to facilitate sharing of methods and components between organisations.

Regarding standard data and metadata formats, SDMX initially seemed to provide the answer. However practical experience in several countries seems to suggest that whilst SDMX is suitable for aggregate data, it is probably not the best solution for microdata flows. Several experiments are currently ongoing in this area[3], but there is a growing view that another standard is needed. The Australian Bureau of Statistics has started to look at a combination of DDI standards for microdata and SDMX for aggregates, within the framework of the GSBPM[4]. Their work is generating a lot of interest from other countries.

The Australian approach also includes the development of a "Generic Statistical Information Model" to complement the GSBPM. A group of national statistical organisations under the name "Statistical Network", led by the Australian Bureau of Statistics, was set up in June 2010 to consider these and other issues relating to shared work towards the industrialisation of statistics.

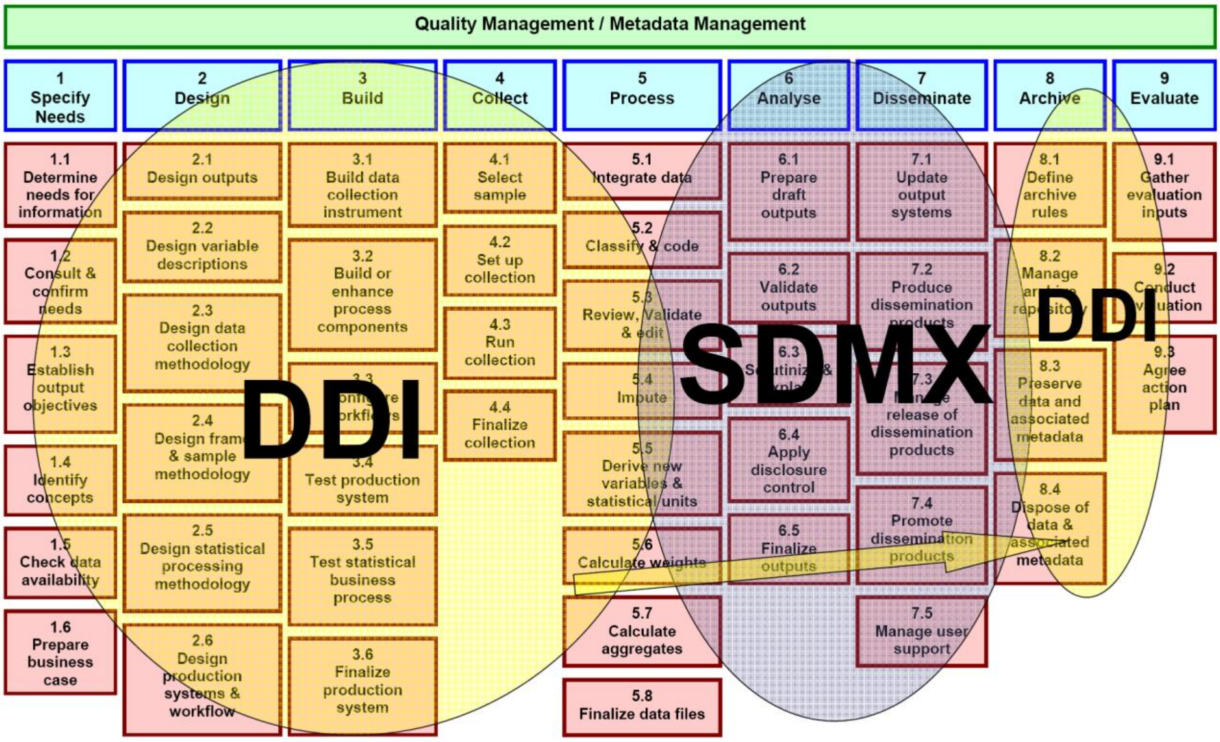
Figure 4 shows a schema of the likely interaction between the GSBPM, DDI and SDMX, based on current thinking.

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

[3] For example in the European Union "ESSnet on SDMX" project: http://sdmxessnet.ine.pt/xportal/xmain?xpid=SDMX&xpgid=sdmx_inst&INST=86019467&xlang=en

[4] See: Studman B. "A Collaborative Development Approach to Agile Statistical Processing Architecture - Australian Bureau of Statistics (ABS) Experience and Aspirations" available at: http://www.unece.org/stats/documents/ece/ces/ge.50/2010/wp.3.e.pdf

Figure 4: Combining the GSBPM, DDI and SDMX


Quality Management / Metadata Management
1 Specify Needs
2 Design
3 Build
4 Collect
5 Process
6 Analyse
7 Disseminate
8 Archive
9 Evaluate
1.1 Determine needs for information
1.2 Consult & confirm needs
1.3 Establish output objectives
1.4 Identify concepts
1.5 Check data availability
1.6 Prepare business case
2.1 Design outputs
2.2 Design variable descriptions
2.3 Design data collection methodology
2.4 Design frame & sample methodology
2.5 Design statistical processing methodology
2.6 Design production systems & workflow
3.1 Build data collection instrument
3.2 Build or enhance process components
3.3 Configure workflows
3.4 Test production system
3.5 Test statistical business process
3.6 Finalize production system
4.1 Select sample
4.2 Set up collection
4.3 Run collection
4.4 Finalize collection
5.1 Integrate data
5.2 Classify & code
5.3 Review, Validate & edit
5.4 Impute
5.5 Derive new variables & statistical units
5.6 Calculate weights
5.7 Calculate aggregates
5.8 Finalize data files
6.1 Prepare draft outputs
6.2 Validate outputs
6.3 Scrutinize & explain
6.4 Apply disclosure control
6.5 Finalize outputs
7.1 Update output systems
7.2 Produce dissemination products
7.3 Manage release of dissemination products
7.4 Promote dissemination products
7.5 Manage user support
8.1 Define archive rules
8.2 Manage archive repository
8.3 Preserve data and associated metadata
8.4 Dispose of data & associated metadata
9.1 Gather evaluation inputs
9.2 Conduct evaluation
9.3 Agree action plan
DDI
SDMX
DDI

All of these developments have attracted the interest of the heads of national statistical organisations, and also in June 2010, a new “High-Level Group for Strategic Developments in Business Architecture in Statistics” was created. This group will oversee developments and work in expert-level groups. It forms part of the work programme of the Conference of European Statisticians, and will report to the annual plenary sessions of the Conference.

4. Conclusions

It is too early to present any firm conclusions on the optimal ways to combine different standards such as DDI, SDMX and the GSBPM. A lot of work still needs to be done before any recommendations can be formulated or best practices identified. The UNECE will continue to encourage this work through the METIS Work Sessions, the new High-Level Group and other relevant international forums.

However, one tentative conclusion is emerging. It is that no single standard is sufficient for the major changes facing the way official statistics are produced. Instead there is an increasing need to apply a number of complementary standards, some already existing, some yet to be developed. The GSBPM, DDI and SDMX are clearly part of the core group of standards, but much more work is needed on how they can best be combined in practice.

5. References

GSBPM

- Documentation and related papers are available at: www.unece.org/stats/gsbpm

DDI

- Specifications and other resources are available at: www.ddialliance.org

SDMX

- Standards and related information are available at: www.sdmx.org