WP.11
ENGLISH ONLY

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Oslo, Norway, 24-26 September 2012)

Topic (ii): Global solutions to editing

Update on the Development of the Generic Statistical Information Model (GSIM)

Prepared by Thérèse Lalor, Australia Bureau of Statistics, Australia
and Steven Vale, UNECE

I. Standards-based modernization

1. Statistical organizations are confronted with shrinking budgets on the one hand and pressure to respond to increasing information needs on the other. The current approach to producing statistics in statistical agencies is characterized by stove-pipe production processes, both at the national and international levels. This lack of integration of processes leads to inefficiencies and often opportunities for common development and sharing of tools, methods and processes are unexplored. Although statistical organizations have the experience and methodology to deal with the data deluge, they do not have the resources to develop the new possibilities.

2. This challenge is being reviewed by the High-Level Group for Strategic Developments in Business Architecture in Statistics (HLG-BAS). The role of this group is to oversee and guide discussions on the modernization of statistical organizations, with a focus on improving the efficiency of the statistical production process, and the ability to produce outputs that better meet user needs.

3. Across the world, statistical agencies undertake activities which largely consume and produce the same information (for example all agencies use classifications, create datasets and publish products). Although the information used by statistical agencies is at its core the same, all agencies tend to describe this information slightly differently (and often in different ways within each agency). There is no common means to describe the information we use. This makes it difficult to communicate clearly within and between statistical agencies and without this there is no foundation for in-depth collaboration and nor greater standardisation and the sharing of tools and methods.

4. HLG-BAS has recognized that there is a need to develop necessary standards to support the modernization of official statistics, including the Generic Statistical Information Model (GSIM).

II. What is GSIM?

5. The GSIM is a reference framework of information objects with generic descriptions of the definition, management, and use of data and metadata throughout the statistical production process. Information objects are entities with agreed names and definitions. They have specified essential properties and relationships with other information objects.

6. GSIM will improve communication at a number of different levels. For example:

- between the different roles in statistical production (statisticians, methodologists and information technology experts);

- between the statistical subject-matter domains;
- between statistical organizations at national and international levels.

7. Improving communication will result in a more efficient exchange of data and metadata within and between statistical organizations, and also with external clients and suppliers.

8. GSIM provides common semantics that can be used unambiguously across and between different implementations. It does not, however, make assumptions about the standards or technologies used in implementation.

III. GSIM Development Project

9. Accelerating the development of the GSIM was agreed as a key priority by the HLG-BAS, following the first HLG-BAS workshop with representatives of expert groups in Geneva (November 2011).

10. In the search for an appropriate mechanism for the next stage of the development of the GSIM, the "sprint" approach (from the "Agile" method of software development) was used. This approach involves gathering a group of experts from all relevant disciplines, and giving them a problem to solve within a specific (and short) time-frame.

11. On this basis, two sprint sessions were approved by the HLG-BAS. The first took place in Slovenia in February, the second in the Republic of Korea in April. Both sprint sessions were two weeks long, and brought together 12-15 experts in statistical production, methodology, information technology, standards and modelling.

12. This approach has significantly accelerated the GSIM development process. In these intensive sessions, it was possible to gain understanding and agreement more quickly than in more traditional methods of collaboration. As a result of the sprints, two new versions of the model were released. Version 0.4 which was released in April 2012 consisted of two documents – an overview document and a communication document. The overview document is aimed at senior management and provides high-level information about the model and its uses and the communication document is aimed at subject-matter staff and provides more detail on the information model.

IV. Specification Layer

13. It was recognized that a further level of detail was needed in order for the model to be implementable. This layer is called the Specification Layer.

14. A major programme of work was required to complete the detailed work involved in the Specification Layer of the GSIM. The work on the Specification Layer occurred between June and August 2012. It involved 28 people from 12 different agencies.

15. For efficiency, four Specification Layer Task Teams, each with a focus on different areas of the GSIM model (Figure 1), were commissioned to undertake this work. A Modelling Team will be created to work with the task teams to develop a consistent and standards aligned formal specification for GSIM. Each team had a weekly two hour meeting to progress the discussions.

16. Of the four areas, the **Conceptual** and **Information** areas have received attention from other collaborations in the past. The work of these teams was influenced by this previous work and also by the relevant standards (for example, DDI and SDMX).

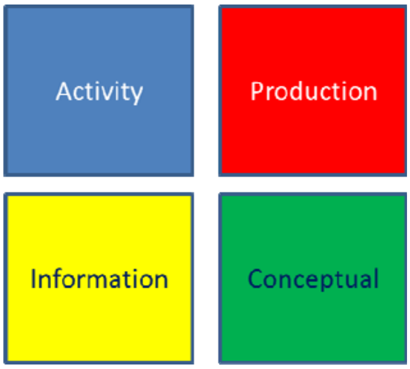


Figure 1. The four areas of work for the Specification Layer.

17. The **Conceptual** team reviewed a family of information objects that describe the concepts used during statistical production processes and their practical implementation. It covers the concepts that allow users to understand what the statistics are measuring. This team considered the information objects surrounding *population*, *variable* and *classification*.

18. The **Information** team reviewed a family of information objects that describe the results of the stages of statistical production. The scope of this team also included a consideration of the information involved in disseminating these results. The information objects described included *dataset, unit data structure definition, cube data structure definition* as well as objects related to dissemination such as *product* (for example a static publication) and *service* (for example a dynamic data service).

19. The remaining two areas, **Activity** and **Production**, were areas which have received much attention in the past. These areas are important to the role of the GSIM in the standards-based modernization effort.

20. The **Activity** team reviewed a family of objects that provide the environmental context in which a set of statistical activities is conducted. This includes objects which occur at the beginning (i.e. business case) and at the end (i.e. evaluation plan) of a statistical production process.

21. This area includes a number of objects related to *Statistical Programme* which is an object that describes the purpose and objectives of a set of activities. *Statistical Programme* will usually correspond to a survey or a survey vehicle, but can cover a related series of surveys, or a single component of a survey.

22. A simple example is where a *Statistical Programme* relates to a single survey, for example, the Labour Force Survey. The *Statistical Programme* will have a series of *Statistical Programme Design* objects. *Statistical Programme Design* objects describe the methodology and design used throughout the life of the survey. When a methodological change is made to the survey, a new *Statistical Programme Design* is created to record the details of the new design. In this way, GSIM provides a mechanism for capturing information about a statistical production process at the highest level.

23. The **Production** team reviewed a family of objects that describe the processes, methods and rules that are used in statistical production. The Generic Statistical Business Process Model (GSBPM) provides descriptions of business processes that can occur throughout the statistical production process. However, in order to describe a process in any level of actionable detail, more information is needed.

24. This additional information is also necessary if you wish to have reusable processes that talk about "flow" rather than just the specific functions which need to be performed during the flow (with no description of how they fit together). As shown in Figure 2, GSBPM processes can be thought of as the vital organs for statistical production and information might be thought of as the blood, nerve signals and hormones.

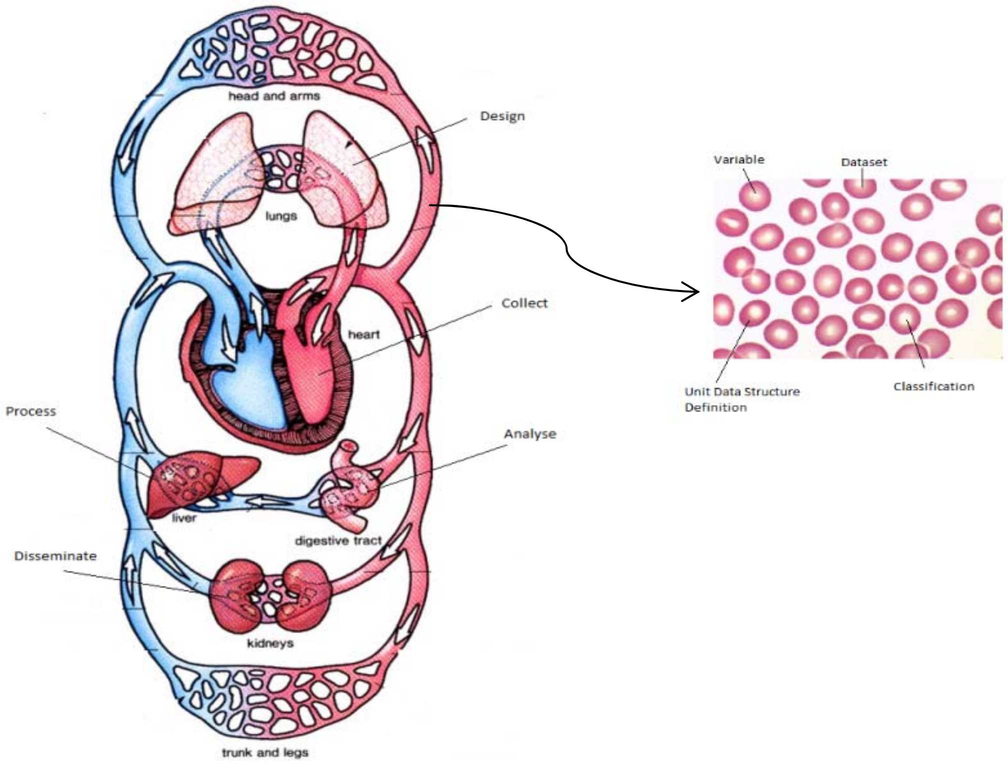


Figure 2. GSBPM and GSIM

25. Information needs to:

- Flow between GSBPM processes. For example, data is processed or transformed between the Collect and Disseminate phases.
- Govern the behaviour of GSBPM sub-processes. There are business rules and derivation formulas that are applied during processes (for example Impute, Derive New Variables). There are also rules or plans that determine which process should be performed next. An example of this is whether the quality of the data is sufficient to proceed to the next step or whether some form of “remedial” processing is required.
- Report on the outcome of GSBPM processes. For example, process-related statistical quality metrics such as response rates or imputation variance.

26. The Production area of GSIM seeks to provide a standard way to capture this information about processes. It includes information objects such as *Process Step Design, Process Step Execution, Business Rule, Process Input* and *Process Output.*

V. Integration Workshop

27. A one-week workshop will be held in The Hague, Netherlands from 17-21 September 2012. The workshop will bring together the members of the task teams who have been working on each area of the Specification Layer as well as members of the overarching Integration Team.

28. The purpose of this meeting is to review the work of each team and form a coherent model. A new version GSIM (v0.8) will be released in the week that follows the Integration Workshop. This will include updated Overview and Communication documents as well as the new Specification Layer document.

VI. Using GSIM

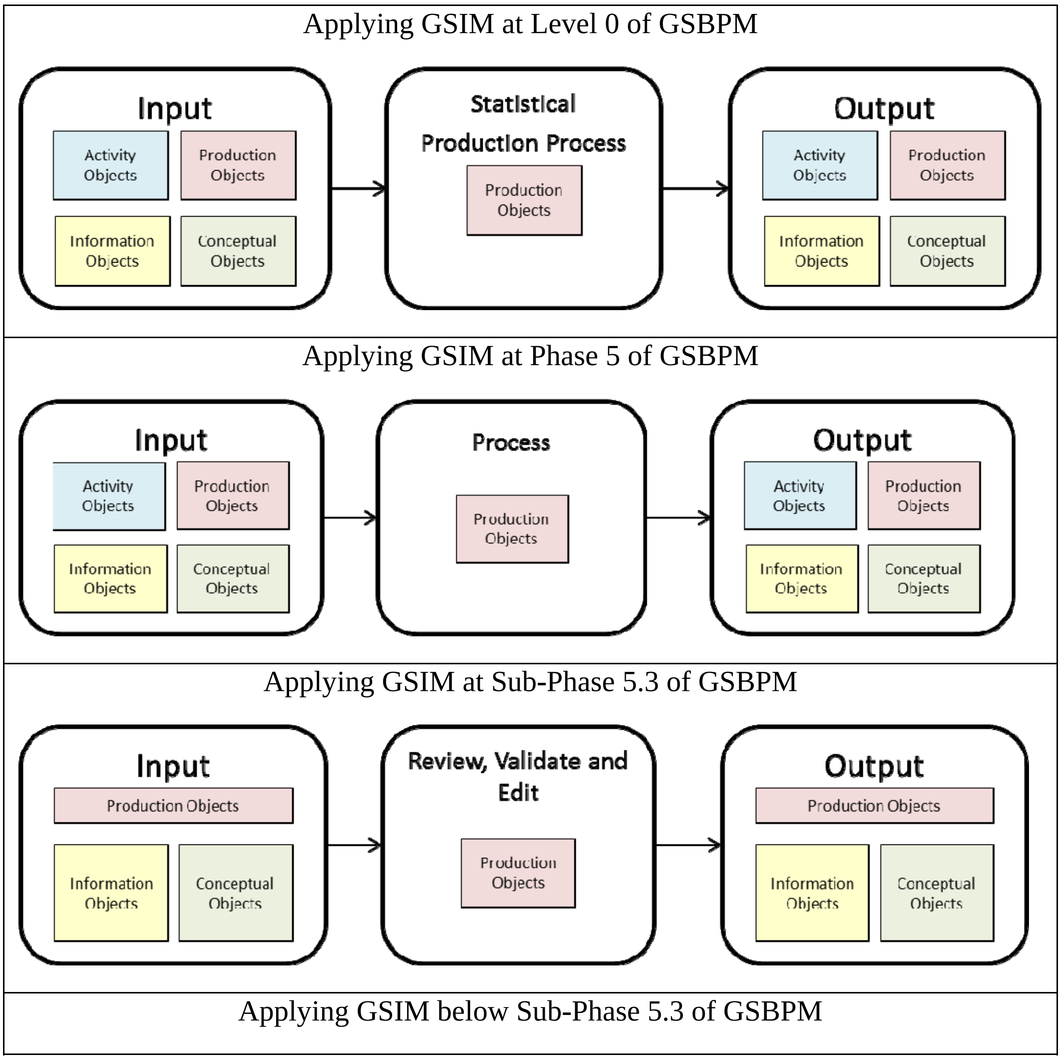
29. In the shorter term GSIM can be used by agencies to:

- Build capability among staff by using GSIM as a teaching aid that provides a simple view of complex information and clear definitions.
- Validate existing information systems and compare with emerging international best practice and where appropriate leverage off international expertise.
- Guide development or update of local or international standards to ensure they meet the broadest needs of the international statistical community.

30. In the longer term, GSIM will support current production processes and facilitate the modernization of statistical production. Implementation of GSIM, in combination with GSBPM, will lead to more important advantages. GSIM will:

- Create an environment prepared for reuse and sharing of methods, components and processes;
- Provide the opportunity to implement rule-based process control, thus minimizing human intervention in the production process;
- Generate economies of scale through development of common tools by the community of statistical organizations.

31. The self-similar principle provides a useful metaphor for describing the creation of statistical production processes. This has the characteristic that if one looks at different levels of magnification one finds that a pattern is repeated at every level, in a recursive fashion. This holds also for the approach for designing statistical production processes using GSIM and GSBPM. At each level we have to design the output and the input, and the process in-between. The output and the input can be designed in terms of information, activity, conceptual and production information objects and the process in-between can be designed using the production information objects.


Applying GSIM at Level 0 of GSBPM

Applying GSIM at Phase 5 of GSBPM

Applying GSIM at Sub-Phase 5.3 of GSBPM

Applying GSIM below Sub-Phase 5.3 of GSBPM

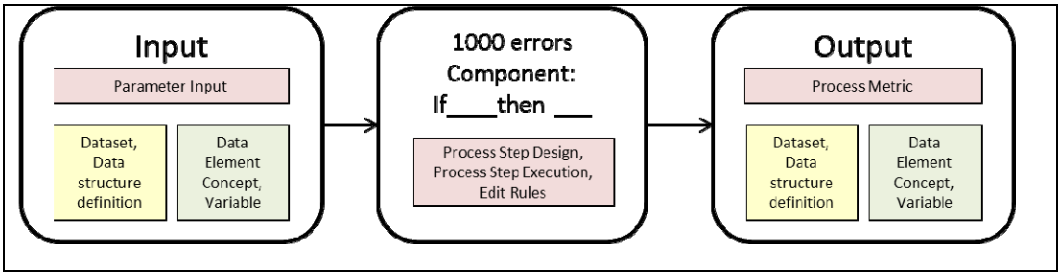

Input
Parameter Input
Dataset, Data structure definition
Data Element Concept, Variable
1000 errors
Component:
If____then ___
Process Step Design, Process Step Execution, Edit Rules
Output
Process Metric
Dataset, Data structure definition
Data Element Concept, Variable

Figure 3. Describing the creation of statistical production processes

32. As seen in Figure 3, the first level can be considered as equivalent to the statistical production process as a whole. The next level corresponds to a phase of the statistical production process (for example column 4, 5, 6 or 7 of the GSBPM). The third level corresponds to a sub-process (for example sub-process 5.3 of the GSBPM – review, validate and edit). The fourth level consists of the individual building blocks within the sub-process. The GSIM facilitates the description of inputs and outputs at each level of the GSBPM, following the same pattern. It provides a consistent structure to design statistical processes.

33. This approach supports the idea of bringing together the GSIM, the GSBPM, harmonized methods and standard technology. It shows how the information, activity, conceptual and production information objects for the GSIM fit together with the structure of the GSBPM when applied in a re-design of statistical production processes. Statistical methods are applied via rules. The components are implemented via technical solutions, and have standard interfaces for inputs and outputs in terms of their structural information objects.

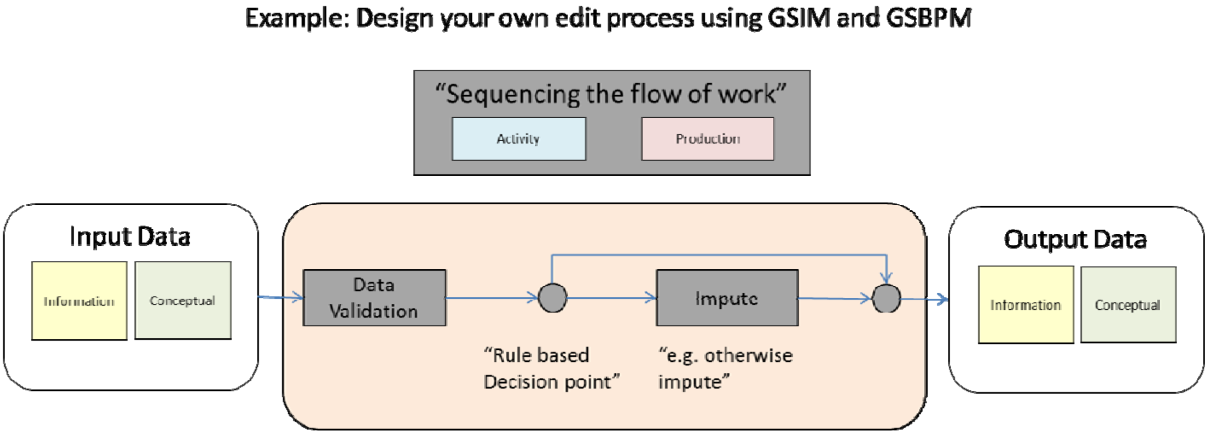


Figure 4. Designing processes using standards

34. It is intended that GSIM may be used by agencies to the degree that they find it of use. It may be used in some cases only as a model to which agencies refer when communicating with other agencies to clarify discussion, in other cases an agency may choose to implement the final GSIM as the information model that defines the agency's operating environment. The modular nature of GSIM means that it is also possible for an agency to choose to use part but not all of GSIM (e.g. information described by the Production group but not that described by the Conceptual group). All these scenarios for the use of GSIM are valid, although those agencies that make use of GSIM to its fullest extent may expect to realize the greatest benefits.

35. A User Guide for GSIM containing further information of how to use GSIM will be released during December 2012.