Generic Statistical Information Model (GSIM):

Communication Paper for a General Statistical Audience

(Version 0.8, September 2012)

DRAFT FOR REVIEW

Please note the development of GSIM is a work in progress. GSIM v0.8 is not intended for official publication.

Instructions for reviewers and a **template for providing feedback** is available at http://www1.unece.org/stat/platform/display/metis/GSIM+v0.8

About this document

This document is aimed at subject matter statisticians, methodologists, process designers, business architects etc. It provides an overview about the information represented in GSIM, and summaries of how the model could be used and relationships to other models and standards.

tion 3.0
e, visit
or part
conomic
national

Introduction

1. Across the world statistical organizations undertake similar activities albeit with variation in the processes each uses. Each of these activities use and produce similar information (for example all agencies use classifications, create data sets and publish products). Although the information used by statistical organizations is at its core the same, all organizations tend to describe this information slightly differently (and often in different ways within each organization). There is no common means to describe the information we use. This makes it difficult to communicate clearly within and between statistical organizations and without this there is no foundation for in-depth collaboration, standardization, or the sharing of tools and methods.

2. The Generic Statistical Information Model (GSIM) is a reference framework of information objects, which enables generic descriptions of the definition, management and use of data and metadata throughout the statistical production process. It provides a set of standardized, consistently described information objects, which are the inputs and outputs in the design and production of statistics. As a reference framework, GSIM helps to explain significant relationships among the entities involved in statistical production, and supports the development of consistent standards or specifications.

3. A model alone cannot transform an organization or its processes, but GSIM is designed to allow for innovative approaches to statistical production to the greatest extent possible; for example, in the area of dissemination, where demands for agility and innovation are increasing. At the same time, GSIM supports more traditional approaches of producing statistics.

4. GSIM is one of the cornerstones for modernizing official statistics and moving away from traditional subject matter silos. By defining objects common to all statistical production, regardless of subject matter, GSIM enables statistical organizations to rethink how their business could be organized to generate economies of scale. It also:

- Improves communication between different disciplines involved in statistical production, within and across statistical organizations; and between users, producers and providers of official statistics.
- Enables configurable, rule-based and modular ways of producing statistics, thus minimizing human intervention in the production process.
- Provides a basis for flexibility and innovation, including support for the easy deployment of new statistical products and the adoption of new types of statistical data sources.

Scope

5. GSIM provides the information object framework supporting all statistical production processes as described in the Generic Statistical Business Process Model (GSBPM), giving the information objects agreed names, defining them, specifying essential properties, and indicating their relationships with other information objects. It does not, however, make assumptions about the standards or technologies used in implementation.

6. The information objects defined include those required for the specification and introduction of new data sources for more innovative data collection, and also the generation of new statistical products.

7. GSIM does not include information objects related to supporting business functions within an organization such as human resources, finance, or legal functions, except to the

extent that this information is used directly in statistical production.

What is an information object?

8. GSIM is an object-oriented model. It contains objects which specify information about the real world – 'information objects'. Examples include data and metadata (such as classifications) as well as the rules and parameters needed for production processes to run (e.g. data editing rules). GSIM identifies around 150 information objects, which are grouped into four broad groups, and are explained in more detail in the specification documentation.

ss Case. a set of **ınel** and terations a given **ːe** is not **vity** will lete, the **ɒurce** to

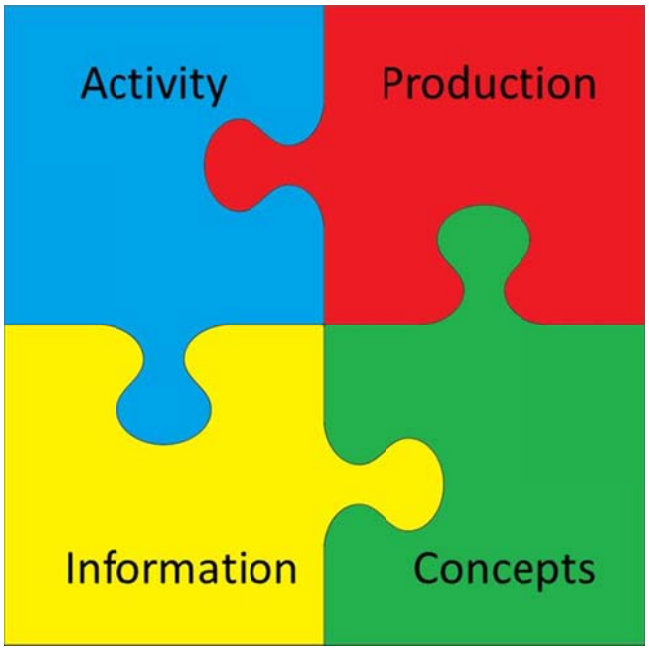


Figure 1: GSIM High-level Information Object Groups

9. Figures 1 and 2 show simplified views of the information objects identified in GSIM. They are not formal models of the entire set of information objects, as that is covered in the more detailed specification document. They can, however, be used as a means for communication with users who are not interested in the detail but are interested in examples of objects and relationships.

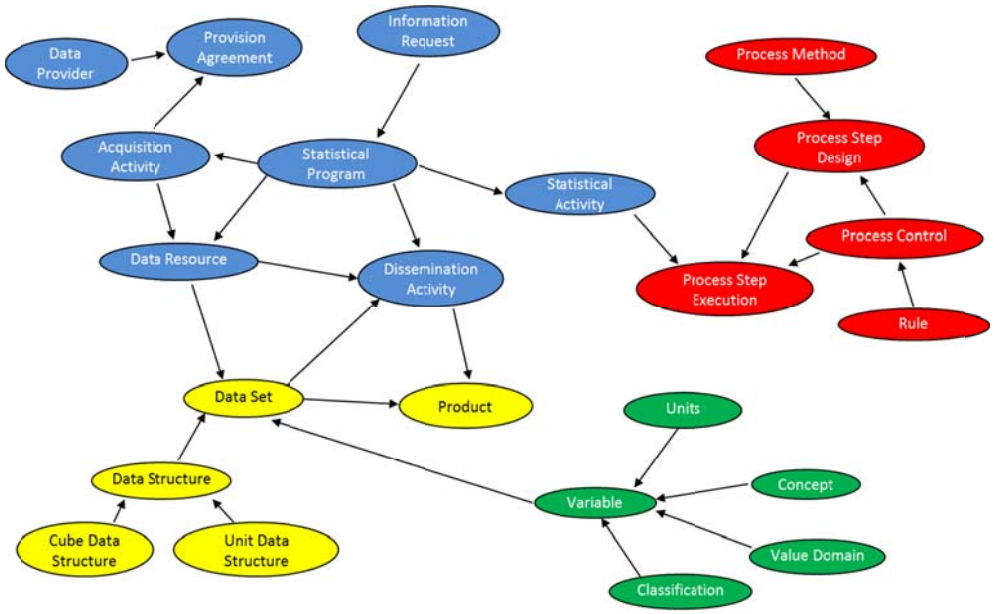


Figure 2: Simplified view of GSIM Information Objects

Background to the development of GSIM

10. The need for GSIM was first discussed at the 2010 Meeting on Management of Statistical Information Systems (MSIS). The work to develop GSIM was given additional impetus in 2011, when the High Level Group for Strategic Developments in Business Architecture in Statistics (HLG-BAS) - a group of heads of national and international statistical organizations that support a common vision to modernize statistical production – identified it as a cornerstone of their vision for the modernization of official statistics.

> *"To enable statistical organizations to arrive at standardized generic industrialized production of statistics, we first need to find one another at the conceptual level....under the umbrella of the GSBPM and the GSIM. This is a very high ambition which will take time. "*
>
> *HLG-BAS Strategic Vision (June 2011)*

11. This resulted in two sprint sessions in early 2012, followed by the establishment of task teams to develop different parts of the model in more detail. An 'Integration Workshop' in September 2012 brought together the different strands to form GSIM version 0.8, which was released for public consultation at the end of that month.

Design principles

12. The following design principles have been established to guide the development and maintenance of GSIM, providing rules to which the model should adhere:

- Principle 0: GSIM has change control i.e. the following principles for designing GSIM apply to every revision of GSIM.
- Principle 1: GSIM supports GSBPM and covers the whole statistical process.
- Principle 2: GSIM can be used independently of any other framework or standard.
- Principle 3: GSIM has an intuitive appeal to all stakeholders.
- Principle 4: GSIM supports the design, documentation, production and maintenance of statistical products.
- Principle 5: GSIM enables explicit separation of the design and production phases.
- Principle 6: GSIM supports both current and new ways of producing statistics.
- Principle 7: GSIM supplies links between process steps at all desired levels of granularity.
- Principle 8: GSIM provides a basis for a common understanding of information objects.
- Principle 9: GSIM uses a layered approach to documentation.
- Principle 10: GSIM contains information objects only down to the level of agreement between key stakeholders.
- Principle 11: GSIM is robust, but can be easily adapted and extended to meet users' needs.
- Principle 12: GSIM information objects and their relationships are presented as simply as possible.
- Principle 13: GSIM makes optimal reuse of existing terms and definitions.
- Principle 14: GSIM does not refer to any specific IT setting or tool.
- Principle 15: GSIM defines and classifies its information objects appropriately, including specification of attributes and relations.

What's new in GSIM v0.8?

13. Version 0.8 of GSIM reflects the work to further develop the detailed model, building on version 0.4 and the feedback received from the public consultation on that version.

14. Key changes between GSIM v0.4 and v0.8 include:

- Addition of the Specification document, setting out the detailed models, definitions and relations that will be necessary for implementations based on GSIM
- Simplification and improvement of the higher-level documentation
- More detail on the relationships between GSIM and a wide range of other standards

15. GSIM uses a layered approach to documentation. The highest level is a two-page overview designed as an introduction to GSIM for non-specialists. The next level comprises this communication paper, also primarily intended for non-specialists. A more detailed specification document has been added in version 0.8, intended for information architects and similar experts, to support the practical implementation of the GSIM approach.

Benefits of GSIM

16. Many statistical organizations are confronted with shrinking budgets and pressure to respond to increasing information needs. Limited integration of processes leads to inefficiencies, both within statistical organizations, and in the international statistical

community. Opportunities for common development and sharing of tools, methods and processes are largely unexplored. Although statistical organizations have the experience and methodology to deal with the data deluge, growing demands and advances in information technology, they do not have the resources to fully explore new possibilities. Statistical production still requires a great deal of manual intervention, which is not only resource intensive, but introduces the potential for human error.

17. A significant benefit of using GSIM as a common language is to improve communication at different levels:

- 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.

18. 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.

19. As a common reference framework for information objects, 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.

20. At a strategic level, GSIM could be used to direct future investment towards areas of statistical production where the common need is greatest. It could also lead to some degree of specialization within the international statistical community. For example, some organizations could specialize in seasonal adjustment, time series analysis or data validation, and other organizations could take advantage of this expertise.

21. In the shorter term GSIM can be used by organizations to:

- Build capability among staff by using GSIM as a teaching aid that provides a simple easy to understand 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 updating of local or international standards to ensure they meet the broadest needs of the international statistical community

22. It is intended that GSIM may be used by organizations to different degrees. It may be used in some cases only as a model to which organizations refer when communicating with other organizations to clarify discussion. In other cases an organization may choose to implement GSIM as the information model that defines their operating environment. Various scenarios for the use of GSIM are valid, although those organizations that make use of GSIM to its fullest extent may expect to realize the greatest benefits.

Methodology, quality and reference metadata in GSIM

Methodology

23. Methodology is reflected in GSIM mainly through rules and parameters. The design of a statistical production process sets out the methodology to be followed when the process is run. The methodology used is therefore likely to be fairly specific either to an individual process or to a group of similar processes.

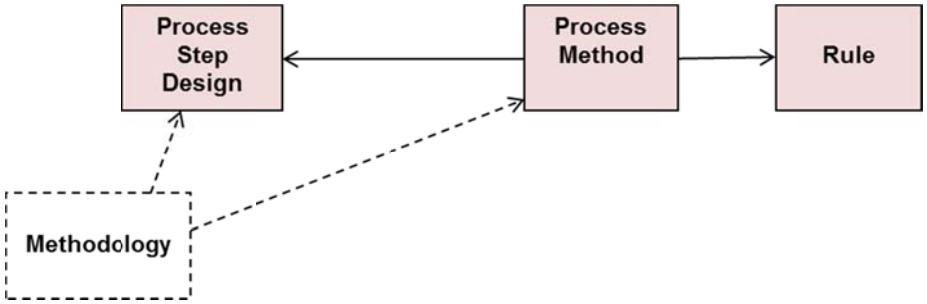


Figure 3: Methodology in GSIM

24. The information object "Methodology" is deliberately not modelled in GSIM, as GSIM is intended to be a generic model, capable of supporting all current and future methods.

ould
f the
r.
ıow'
f the
ın of
step
: has

iable

then
rror'

Quality

25. While methodology is embedded in the design of the statistical production process, quality is linked to the instance (i.e. to production runs) of the process. Quality itself can have many forms depending on the purpose that it is there for, such as the organizational aspect, the quality of the processes or the quality of the statistics. Quality reports traditionally mainly refer to the quality of the statistics. Quality is relevant at a number of different levels of instances of information objects. For example, as an attribute to an information element (e.g. quality flag), as an attribute to a data set (e.g. status provisional data, final data, revised data). It also appears as process quality information. The product quality is laid down in a quality report, which is itself also a statistical product.

26. Quality information and quality reports can be tied to the production process as a whole and/or to parts of it. Quality is present in the inputs and outputs of process steps in GSBPM, acting as process data to control and steer rules for processes. Quality can be an output of the information objects in GSIM combined with the processes in GSBPM.

27. Quality therefore means different things in different settings. Depending on the scope, it will refer to different information objects in relation to relevant processes. In order to have Quality as an information object in GSIM we have to establish what we mean by it. Quality is therefore not seen as an explicit object in GSIM, at least for version 0.8.

28. Quality frameworks are treated in a similar way to methods, as there is as yet no single generic international quality framework.

Reference metadata

29. In a similar way to Quality and Methodology, reference metadata is seen more as an attribute of information objects, rather than an object in its own right. For example, conceptual metadata are represented in GSIM in the group of conceptual objects, methodological and procedural aspects of metadata are represented by GSIM objects in the production group. Other aspects of metadata may be modelled by means of object attributes. Reference metadata can be attached to any information object and, as such, can be an input to as well as an output of a GSBPM process step.

30. One reason for this approach is that there is currently no globally agreed definition of the scope of reference metadata, and the GSIM should remain generic rather than linked to one specific definition. As for Methodology, GSIM can, of course, be extended during implementation to fit specific reference metadata models and requirements.

Relationship to GSBPM

31. GSIM and GSBPM are complementary models for the production and management of statistical information. GSBPM models the statistical production process and identifies the activities undertaken by producers of official statistics that result in information outputs. These activities are broken down into sub-processes, such as “Impute” and “Calculate aggregates”. GSIM helps describe interrelated sub-processes by defining the information objects that flow between them, that are created in them, and that are used by them to produce official statistics.

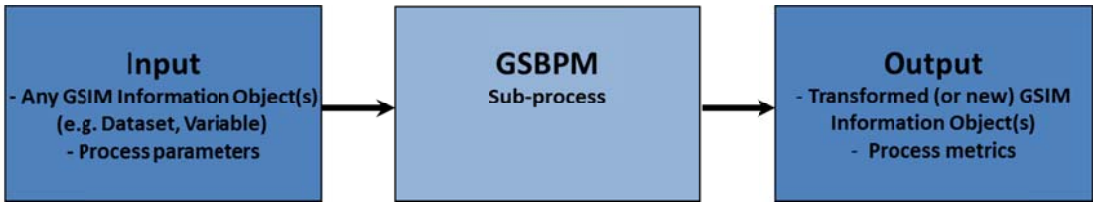


Figure 4: GSIM and GSBPM

32. As described in the strategic vision of the HLG-BAS, much greater value will be obtained from GSIM if it is applied in conjunction with GSBPM. Likewise, greater value will be obtained from GSBPM if it is applied in conjunction with GSIM. Nevertheless, just as GSBPM has been applied to date without GSIM, it is possible (although not ideal) to apply GSIM without GSBPM. For example, an agency may currently be using a local variation on GSBPM to model their statistical business processes, rather than using GSBPM itself. This decision in regard to modelling statistical business processes should not necessarily prevent them deciding to apply GSIM as a reference framework for statistical information.

33. In the same way that individual statistical business processes do not use all of the sub-processes described within GSBPM, not every information object in GSIM is necessarily required to be used and/or produced in the course of every statistical business process.

34. Applying GSIM together with GSBPM (or an organization-specific equivalent) can:

- facilitate building efficient metadata driven collection, processing, and dissemination systems.
- help harmonize statistical computing infrastructures.

35. Good metadata management is essential for the efficient operation of statistical business processes. Metadata are present in every phase of GSBPM, either created, updated or carried forward unchanged from a previous phase. In the context of GSBPM, the emphasis of the over-arching process of metadata management is on the creation, updating, use and reuse of statistical metadata. Metadata management strategies and systems are therefore vital to the operation of GSBPM, and are facilitated by GSIM.

36. GSIM supports a consistent approach to metadata, facilitating the primary role for metadata envisaged in Part A of the Common Metadata Framework "Statistical Metadata in a Corporate Context", i.e. that metadata should *uniquely and formally define the content and links between objects and processes in the statistical information system*.

Relationships to other standards

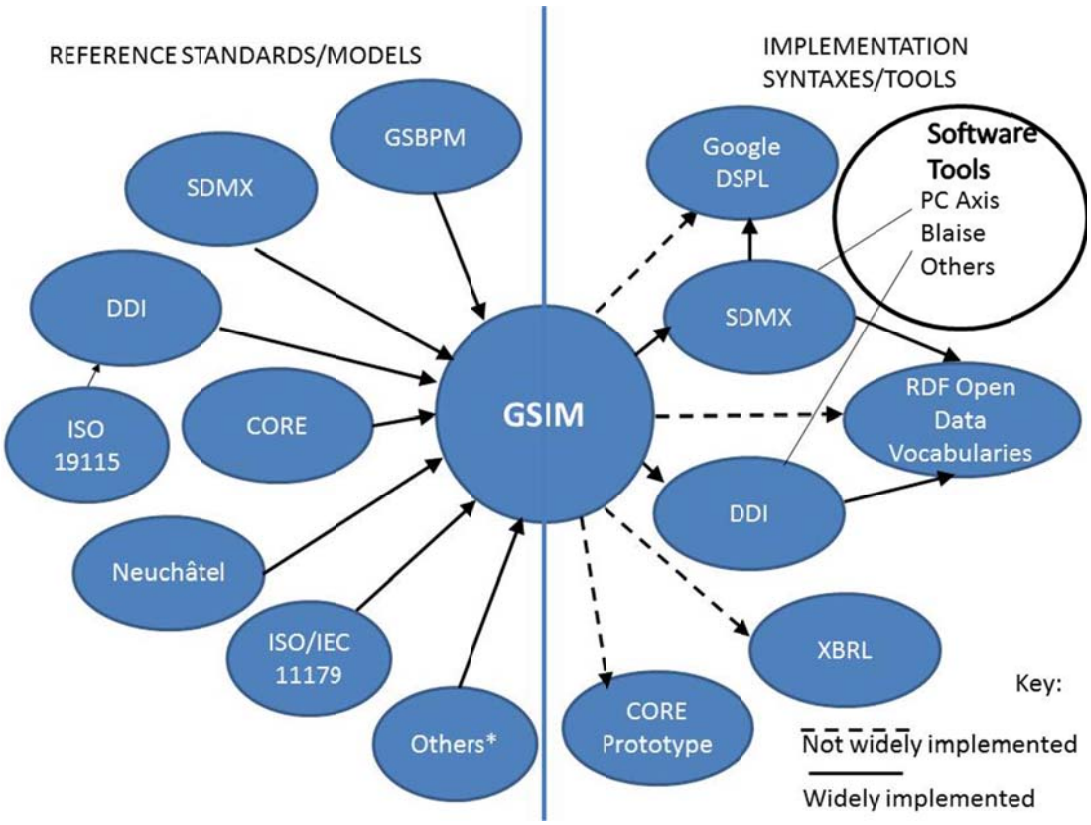
37. One of the design principles of GSIM is to make optimal reuse of existing terms and definitions, wherever possible, to facilitate the use of the reference framework. In developing GSIM, many existing models and standards have been examined, both to determine the best approach for describing GSIM objects, and also to test the completeness and usability of the model. The specification layer of documentation explores relationships between GSIM and a number of other standards and models in detail.

38. GSIM must be implementable. In order to support the implementation of GSIM, known standards and tools have also been examined, to ensure that the reference framework is complete and useful in this respect. The relationship between GSIM and other models and

standards is two-fold. The standards and models serve as inputs to the creation of GSIM, and also act as targets for the use of GSIM within organizations.

39. By taking this approach, it is hoped that GSIM will be as similar as possible to the information which user organizations already have within their statistical production systems, allowing GSIM to be more understandable and easier to implement.

40. Figure 5 illustrates how different relevant standards, models, and implementation syntaxes and tools relate to GSIM. Standards and models that have provided significant input to GSIM are presented on the left hand side of the figure. Implementation syntaxes and tools that are currently of relevance to an implementation of GSIM are presented on the right hand side of the figure. This list will become outdated as more and more implementation syntaxes and tools are developed. The particular software packages listed are widely used in statistical organizations, but are intended to be illustrative examples, and are not a complete list.



** There are too many others to show in the diagram*

Figure 5: GSIM and its relationship to other relevant standards and models.

Future work on GSIM

Roadmap for getting to GSIM v1.0

41. A detailed roadmap, outlining the work to be done to develop public release v1.0 of GSIM has been prepared. The remaining steps are:

- Collection of feedback on GSIM v0.8
- Discussion at the Workshop on Strategic Developments in Business Architecture in Statistics (7-8 November 2012)
- Drafting of GSIM v1.0 - An accompanying Communication Plan and User Guide will also be prepared.

Beyond GSIM v1.0

42. As a newly developed framework, it is expected that additions and changes will be identified as GSIM is applied in practice. An updated version of GSIM (e.g. v1.1) may be warranted, for example, within a year of the release of v1.0.

43. Processes will be established to capture feedback from practical use and feed this into further evolution of GSIM. A process for setting and developing a release schedule will be established, together with a process for design and stakeholder review of proposed new releases. New releases will be designed in a way which minimizes the extent of change from previous versions (in order to maximize backwards compatibility) while still meeting the business needs which initiated development of a new release.

44. Release of updated versions would be expected to become less frequent once the initial set of additional requirements encountered through widespread practical application of GSIM has been addressed. New requirements within the community of producers of official statistics, however, could still initiate the need for updated versions of GSIM.