United Nations ECE/CES/2024/3

Economic and Social Council

Distr.: General
9 April 2024

Original: English

Economic Commission for Europe

Conference of European Statisticians

Seventy-second plenary session
Geneva, 20 and 21 June 2024
Item 10 (c) of the provisional agenda
Reports, guidelines and recommendations prepared under the umbrella of the Conference: New version of the Generic Statistical Information Model

New version of the Generic Statistical Information Model

Prepared by the Secretariat

Summary

The Generic Statistical Information Model (GSIM) describes and defines a set of information used in a statistical production process. By standardizing information that flows within the production process, GSIM can facilitate reusing and sharing of methods, tools and processes within and between statistical organizations.

This note provides a brief overview of GSIM, the review process carried out under the High-Level Group for the Modernisation of Official Statistics (HLG-MOS) Modernisation Group on Supporting Standards and changes in the new version of the model. The previous version of the model was endorsed by the Conference of European Statisticians (CES) in 2020.

The Conference is invited to endorse the new version of GSIM, which is presented in document INF.2 in a printable form and on this GitHub site.

I. Background

1. Across the world, statistical organizations undertake similar activities, albeit with variations in the processes that they use. Each of those activities uses and produces similar information (for example, all organizations define populations for their statistical observations, use statistical classifications, create data sets and disseminate information). Although the information used by statistical organizations is at its core the same, all organizations tend to describe it slightly differently and sometimes in different ways within each organization.

2. The Generic Statistical Information Model (GSIM) is a reference framework for statistical information, first developed in 2012. It provides a set of standardized, consistently described information classes, which can be used as inputs and outputs in the design and the production of statistics. As a reference framework, GSIM explains significant relationships among the entities involved in statistical production, and can be used to guide the development and use of consistent implementation standards or specifications.

3. As a common language to describe statistical information, GSIM can facilitate communication within and between statistical organizations. It can provide the foundation for in-depth collaboration, standardization, or sharing of tools and methods, and thereby plays an important role in modernizing, streamlining and, aligning standards and production associated with official statistics at both national and international levels. GSIM is one of the cornerstones for modernizing official statistics and moving away from subject matter silos. It is a key element of the strategic vision of the High-Level Group for the Modernisation of Official Statistics (HLG-MOS).[1]

4. The modernization of statistical production is needed for statistical organizations to remain relevant and flexible in a dynamic and competitive information environment. It is hoped that statistical organizations will adopt and implement GSIM and the common language it provides.

5. Section II of this document gives an overview of GSIM, while Section III describes the progression of its development. Section IV then outlines main changes made in the new version of the model, and Section V gives a brief overview of the review process. Lastly, Section VI discusses the benefits for an organization to implement GSIM.

II. Overview of the Generic Statistical Information Model

A. Scope

6. In terms of its scope, GSIM covers statistical business processes as described in the Generic Statistical Business Process Model (GSBPM),[2] giving the information classes agreed names, defining them, specifying their essential properties, as well as indicating their relationships with other information classes. It does not, however, make assumptions about the standards or technologies used to implement the model.

7. GSIM does not include information classes related to activities within an organization such as human resources, finance, or legal functions, except to the extent that this information is used directly in statistical production. For more information on these activities, see the Generic Activity Model for Statistical Organisations (GAMSO).[3]

8. GSIM is a conceptual model. Organizations can choose existing "implementation standards", such as Statistical Data and Metadata eXchange (SDMX) and Data Documentation Initiative (DDI) for the technical implementation of standards in production processes. Rather, GSIM provides a common reference point at the conceptual level with which to compare and contextualize implementation standards.

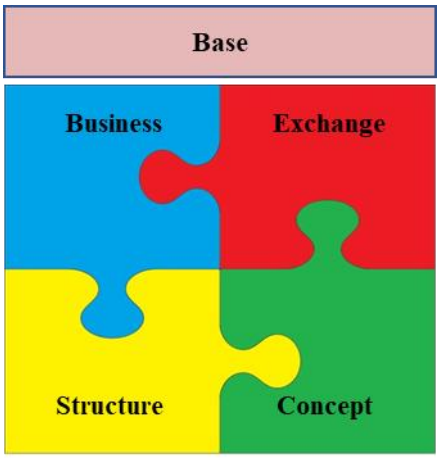
[1] UNECE Statistics Wikis – HLG-MOS (https://statswiki.unece.org/display/hlgbas).
[2] UNECE Statistics Wikis – GSBPM (https://statswiki.unece.org/display/GSBPM).
[3] UNECE Statistics Wikis – GAMSO (https://statswiki.unece.org/display/GAMSO).

B. Main elements

9. The GSIM model is comprised of specific classes that define information about the real world and can be described as “information classes”. GSIM consists of nearly 130 information classes which are pieces of information needed for a statistical production process (such as “Conceptual Variable”, “Statistical Classification”, “Process Input”). In the previous versions of GSIM, they were referred to as “information objects”.

10. For simplicity, GSIM groups these various classes into five broad types based on their role in statistical production, as shown in the following Figure:

GSIM Top-level information class groups



11. The five GSIM groups can be summarized as follows:

(a) Base group that provides features that are reusable by other information classes to support functionality such as identity and versioning;

(b) Business group that is used to capture the design and execution of statistical programmes and supporting activities;

(c) Concept group that is used to define the semantics and representations of data;

(d) Exchange group that is used to model the sharing of data and metadata, both internally and externally; and

(e) Structures group that is used to structure information throughout the statistical process.

12. Each information class has a definition, explanatory text, and a list of attributes. These definitions, along with the relationships among classes reflect how statistical organizations use information during the production process.

III. Progression of the Generic Statistical Information Model’s development

13. The model was first developed in 2012 and subsequently revised several times to reflect user experiences and the changing business environment. The model was endorsed by CES in 2017 along with GSBPM, GAMSO, and the Common Statistical Production Architecture (CSPA). The previous version of the model (version 1.2) was released in 2019 and endorsed by CES in 2020.

14. In 2020, the HLG-MOS Modernisation Group on Supporting Standards (further referred to as the Supporting Standards Group) set up a task team for additional minor updates to GSIM. Minor updates primarily refer to the addition of information classes or the change of a class’s name that can be mapped one-to-one to its previous version, the re-wording of

definitions without changing the core meaning, or the correction of typos.[4] However, works being carried out by other Supporting Standards Group task teams, such as the "Linking GSBPM and GSIM" task team, prompted discussions that GSIM would require more than minor updates to improve consistency and user-friendliness, especially for GSBPM users. Consequently, the Supporting Standards Group requested HLG-MOS to extend the mandate of the GSIM task team from a minor update to preparing a full new version. The formal process therefore took place in 2022–2023.

15. That task team was chaired by Canada, and consisted of the following members: Egypt, Finland, France, Hungary, Italy, Mexico, the Netherlands, Norway, the United Kingdom of Great Britain and Northern Ireland, the United States of America and the United Nations Economic Commission for Europe (UNECE). The Supporting Standard Group oversaw the work of the task team to ensure the new version is aligned with other HLG-MOS models such as GSBPM and GAMSO.

16. In keeping with the aim of using modern approaches to foster collaboration in official statistics, the new version of GSIM (version 2.0) was released in a dedicated GitHub repository, which has become a very popular means of releasing resources in recent years. It is anticipated that housing the model in an open platform such as GitHub can maximize its adoption and leverage the potential for communities of users to suggest enhancements to GSIM that could be considered in its future versions.

IV. Main changes since version 1.2

17. Compared to the earlier version 1.2 of GSIM, the main changes included:

(a) Six information classes were added (e.g., "*Register*", "*Reference Document*", "*Software Agent*") and eight information classes were removed (e.g., "*Administrative Register*", "*Statistical Register*", "*Dimensional Data Point*", "*Unit Data Point*") to simplify the model and to reflect changes in how statistical organizations work;

(b) Five information classes were re-named (e.g., "*Transformable Input*" to "*Core Input*", and "*Exchange Channel*" to "*Exchange Instrument*") to be applicable for various types of data (e.g., non-transformable inputs such as registers) in the statistical production and to clarify their meaning;

(c) Relationships among information classes were updated, with major updates being around classes related to referential metadata based on real use cases such as quality reporting. Another area of focus was information classes related to designing of business processes and exchange instruments, to better separate what happens during design and implementation.

18. More details about changes made can be found in the summary page and the change log in the GitHub repository.

V. Review process

19. The GSIM review involved two rounds of consultations: in 2022 and 2023.

20. The first round was made by the GSIM task team, based on feedback from GSIM users (collected in March 2022) as well as from other task teams under the Supporting Standards Group, such as on linking GSBPM and GSIM (2021).

21. The new version of the model was sent for review to all CES member countries in June-July 2023. Feedback was received from Belarus, Canada, Croatia, Ecuador, Finland, Malta, Mexico, and Norway either by email or via a GitHub consultation repository, and analysed by the Secretariat in consultation with leading task team members, resulting in a

[4] ModernStats Models Governance Guidance https://statswiki.unece.org/display/hlgbas/HLG-MOS+Outputs?preview=/120128748/338329756/ModernStats%20Model%20Governance.pdf.

small number of further refinements to the model, allowing it to be released on GitHub in December 2023.

22. As mentioned earlier, the implemented changes were released in the GitHub repository in the form of definitions, explanatory text and attributes (which are also documented in the addendum to this document). That repository also contains interactive diagrams that describe relationships between the information classes, which users can navigate through by clicking on a given class to view its relationships with its nearest neighbours in the model.

23. The following table summarizes the main stages of the GSIM review process.

March 2022	Initial feedback collected (supplemented by feedback previously received from prior Supporting Stands Group works)
June 2023	CES-wide consultation on a proposed new version of the model
December 2023	GSIM 2.0 released on GitHub

24. In addition to the new version of the model, the task team also worked on a "GSIM User Guide" based on the GSIM Specification (2013) and the update of GSIM Communication Paper (2020). These materials will also be made available in the GitHub repository.

VI. Benefits of using the Generic Statistical Information Model

25. It is envisaged that GSIM may be used by organizations to different extents. It may be used in some cases only as a model to which organizations refer when communicating to clarify the discussion, either internally or with other organizations. In other cases, an organization may choose to implement GSIM as the information model that defines its 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.

26. A significant and immediate benefit of using GSIM is that it provides a common language to improve communication at different levels:

- Between the different roles in a statistical business process (business and information technology experts)
- Between the different statistical subject matter domains
- Between statistical organizations at national and international levels.

27. Improving communication can result in a more efficient exchange of data and metadata within and between statistical organizations, and with external information providers and consumers.

28. In the longer term, implementation of GSIM, in combination with GSBPM, could lead to benefits such as:

- Creating an environment prepared for reuse and sharing of methods, code and processes
- Providing the opportunity to automate processes, thus minimizing human intervention in statistical production
- Facilitating the generation of economies of scale through the development of common tools by the community of statistical organizations.

29. GSIM could be used to direct future investment towards areas of statistical production where the common need is greatest. It could also enable some degree of specialization within the international statistical community, allowing some organizations to specialize in

particular areas and to share their tools and expertise to avoid duplication of efforts by other organizations.

30. By defining standardized information classes common to all statistical production, regardless of subject matter, GSIM enables statistical organizations to rethink how their business could be more efficiently organized. By providing common context and linkage points, GSIM and GSBPM can help to make decisions about when and where to use implementation standards such as SDMX and DDI with a view to maximizing automation and interoperability. Such considerations are relevant to those wishing to construct data pipelines.

31. It should also be noted that GSIM and other conceptual models are relevant to the emerging field of Artificial Intelligence (AI). While recent advances in AI offer a significant opportunity for statistical organizations to make their operations more efficient, AI alone cannot transform the statistical production process.

32. In the context of AI, standards can play a pivotal role in structuring processes and improving the quality of information (e.g., enriching data with semantics in the form of machine-actionable standardized metadata), leading to operational efficiencies and automation.

33. Standards such as GSIM therefore provide a foundation upon which AI-assisted production can fully leverage its potential. Since standards improve the qualities of outputs generally, it is expected that AI models would benefit from this, for example, by improving the accuracy of results, reducing the guesswork needed for AI to interpret data, or by making the basis of generated results somewhat more transparent.