Implementing GSIM

Generic Statistical Information Model (GSIM): Implementing GSIM

(Version 1.1, December 2013)

About this document
The document contains information to help users understand what a conceptual model is. It provides guides to help understand GSIM, explains how GSIM can be used as a communication tool and describes the steps to be taken to implement GSIM at a business and technical level.

This document should be read in conjunction with either the GSIM Communication paper or the GSIM Specification.

This work is licensed under the Creative Commons Attribution 3.0 Unported License. To view a copy of this license, visit http://creativecommons.org/licenses/by/3.0/. If you re-use all or part of this work, please attribute it to the United Nations Economic Commission for Europe (UNECE), on behalf of the international statistical community.

Table of Contents

Introduction
A conceptual model
Understanding GSIM
GSIM as a communication tool
Implementing GSIM at a business level
Implementing GSIM at a technical level
DDI Profiles for Use with GSIM Implementations
What is a DDI Profile?
Common Statistical Production Architecture (CSPA)

Introduction

1. GSIM V1.0, released in December 2012, was the first internationally endorsed reference framework for statistical information. This over-arching framework will play an important part in modernizing, streamlining and aligning the production and standards associated with official statistics at both national and international levels.

2. GSIM V1.0 was the result of extensive, multi-disciplinary development and consultation across the international statistical community. Based on initial implementation experiences and further reflections on specific aspects of the model, GSIM V1.1 was released in December 2013. Over time it is expected that GSIM will be further refined as a result of implementation experience.
3. It is expected that statistical organizations will progressively adopt and implement this reference framework and the ‘common language’ it provides. It is intended that GSIM may be used by organizations to different degrees.

4. The document contains information to help users understand what a conceptual model is. It provides guides to help understand the model, explains how GSIM can be used as a communication tool and describes the steps to be taken to implement GSIM at a business and technical level.

A conceptual model

5. GSIM is a conceptual model. Being a conceptual model, the focus of GSIM is on high-level concepts and the relationships between them, and not on implementation details.

6. By design, GSIM does not refer to any specific IT setting or tool. Statistical organizations use a wide range of in-house and proprietary hardware and software platforms; this environment also changes over time. GSIM is designed to be platform-independent in order to be relevant to all stakeholders and robust over time.

7. 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 organizations use 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 this information slightly differently (and often in different ways within each organization). Before GSIM, there was no common way to describe the information we use.

8. GSIM defines and describes the pieces of information (called information objects) that are important to statistical organizations. It also gives users the relationships between the information objects. By describing statistical information in a consistent way, statistical organizations become able to communicate unequivocally and to collaborate more closely (both at a national and international level).

9. A conceptual model such as GSIM cannot be implemented directly. It requires a further level of detail. The sections on Implementing GSIM at a business level and Implementing GSIM at a technical level provide more detail on how this might be done.

Understanding GSIM

10. The documentation of GSIM v1.1 has a number of layers. Each layer provides more detail on the model and is targeted at different audiences. There is:

   - A number of brochures which provide a high level introduction to GSIM
   - A communication document which gives an overview of the model. It also describes the scope and benefits, what the model means for staff in a statistical organization and relationships to other standards and models.
   - Implementing GSIM (this document), which explains how GSIM can be used and implemented.
   - Clickable GSIM – a tool on the UNECE wiki platform that allows users to easily navigate through different views of GSIM to investigate the objects that interest them.
   - A specification document, which provides deeper detail about the model, and an explanation of the GSIM Extension Methodology.
   - An Enterprise Architect file, which contains the UML descriptions of the model.

11. To help target the efforts of staff in statistical organizations, Table 1 provides some guidance on which documents are most appropriate for different audiences.

<table>
<thead>
<tr>
<th>Audience</th>
<th>Suggested documents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top &amp; Senior managers</td>
<td>GSIM Brochures</td>
</tr>
<tr>
<td>Middle Managers, Subject Matter Statisticians</td>
<td>GSIM Brochures, GSIM Communication document, Implementing GSIM, Clickable GSIM</td>
</tr>
<tr>
<td>Architects, business analysts and metadata</td>
<td>GSIM Communication document, Implementing GSIM, Clickable GSIM, GSIM Specification</td>
</tr>
<tr>
<td>Solution Architects</td>
<td>Implementing GSIM, Clickable GSIM, GSIM Specification, Enterprise Architect file</td>
</tr>
</tbody>
</table>

12. A major barrier to effective collaboration within and between statistical organizations has been the lack of common terminology.

13. A ‘survey’ in Statistics Canada, for example, is a ‘survey instance’ according to the UNECE, a ‘collection cycle’ according to Australian Bureau of Statistics, and a ‘study’ according to the external research community. These examples are just the tip of the iceberg.
14. This has made it difficult to communicate clearly within and between statistical organizations and without a common statistical language, there is no foundation for in-depth collaboration, standardization, or sharing of tools and methods.

15. GSIM, as the unifying ‘common language’ for official statistics, will enable rationalization, cooperation and collaboration within and between statistical organizations.

**Box 1. Other industries use information models**

The Official Statistics industry is not alone in recognizing the need for a ‘common language’ to underpin standards-based modernization.

Travellers, for example, can make their own arrangements to travel anywhere in the world by selecting and booking available flights. Their travel is enabled by a standardized aviation information model, which support streamlined flight planning and in-flight navigation. These models also support air-traffic control to reliably handle increased traffic, reduce fuel use and coordinate on-time arrivals and on-ground services. These standardized information models have promoted the adoption of an industry-standard ‘common language’.

GSIM, as an information model, provides the equivalent common, unifying language for the official statistics industry.

16. GSIM provides staff with simple, easy to understand views of complex information. By describing statistical information in a consistent way, statistical organizations become able to communicate unequivocally and collaborate more closely.

17. Information can be a vague concept. Often when staff are asked to describe the information that is the input to and output from a statistical process, this can be a difficult task. It is difficult because most people don’t know where to start.

18. GSIM can be used to educate staff. It gives them a framework for thinking about information. It provides a common terminology to describe the processes and their inputs and outputs that are used to generate official statistics. At the simplest level, staff could look at Figure 2 (below) and start to identify whether these information objects are relevant to them.

19. The layered documentation of GSIM means that staff can find the appropriate level of detail for them. For some, this level is an overview of the information objects. Others will want to see all the information objects, their attributes and relationships.

**Figure 2. Simplified view of GSIM information objects**

20. By providing these simple views on statistical information, GSIM can be used as a way to help staff understand what technical standards such as Data Documentation Initiative (DDI) and Statistical Data and Metadata eXchange (SDMX) describe. If staff can relate these standards to more familiar terms, they become easier to understand and use. For example, knowing that in DDI a ‘Universe’ is the same as the Population information object in GSIM helps.

21. Using GSIM as a common language will increase the ability to compare within and between statistical organizations. All processes that lead to the production of statistics can be described in this one integrated model. This includes the analysis of business needs, the establishment of statistical programs, the development and management of statistical methods, the design of production processes and their cyclical execution.

22. For example, GSIM can be used by staff (at all levels) to share and compare the concepts used in their work. It is also agnostic of subject matter, so it can be used to compare statistical production across subject matter departments.
Implementing GSIM at a business level

23. GSIM can either be mapped to an existing information model or adopted as is by a statistical organization. The steps to adopting GSIM are outlined in the following two sections. Please see the GSIM Specification for further details about the information objects mentioned in these sections.

A statistical organization has an existing information model

24. It should be a straightforward task to map an existing information model onto GSIM. The order in which the mapping is undertaken could be dependent on the orientation of the existing information model.

25. If the information model focuses primarily on metadata, start with the Concepts group of the GSIM. This group comprises information objects such as Variable, Statistical Classification, and Value Domain.

26. If the information model stresses mainly the management of data sets, the point of departure could be the Structures group. This is where notions such as Data Set, Referential Metadata Set and information about their structures can be found.

27. The information model could be principally orientated on the high level management of the statistical process as well as the design and execution of statistical processes. In this case, start the mapping with the Business group. It offers concepts such as Statistical Need, Assessment, Business Case, Statistical Program, Process Step Design, Input Specification and Output Specification.

28. It might, however, be preferable to begin with information objects referring to collection and dissemination of information. The Exchange group includes information objects referring to Questionnaire, Administrative Register, Product, Information Provider and Information Consumer.

29. These four groups are interrelated through relations linking information objects across the group borders, so you will easily find a path from one group to the other.

A statistical organization adopts GSIM as its information model

30. Most statistical organizations have some information models distributed over one or more repositories (catalogues, data bases, etc.) to manage statistical methods, statistical metadata, architectural principles, policy provisions and similar things. In many cases these information models may be implicit rather than explicit. Typically these models will represent a subset of the information objects in GSIM – containing just those that are relevant to the purpose of the particular model.

31. A statistical organization may choose to:

- adopt one of the existing models as the organization's preferred information model, map this model to GSIM and adopt the sections of GSIM which are needed to address any gaps in coverage; or
- adopt GSIM to bring all this information into one consistent model.

32. The order in which the different bits of information are brought together under the GSIM will be dependent on the relative importance of the collections of information.

GSIM and GSBPM

33. Although GSIM can be used independently, it has been designed to work in conjunction with the Generic Statistical Business Process Model (GSBPM). It supports GSBPM and covers the whole statistical process. It is assumed in this section that an organization either uses GSBPM or uses another business process model (which can be mapped to GSBPM).

34. Adopting GSIM at a business level involves an analysis of the information being used, managed and processed when designing and producing statistics. When designing a new process or redesigning an existing process, the process should be mapped to GSBPM and the information objects should be mapped to GSIM. In Annex A, there are a number of examples of how this can be done.

35. This work is useful, because GSIM makes business processes and methods "visible", where other design approaches keep them buried in application code and documentation. This opens up a range of technical possibilities. For example:

- Comparing IT solutions
- Sharing IT solutions across subject matter domains, or even between statistical organizations

36. After undertaking this exercise, it is possible that there will be information objects that your organization needs to describe, but that are not accounted for in GSIM.

37. Most organizations have legacy systems and administrative practices that will require an extension of GSIM to meet organization-specific implementation needs. In particular, processes relating to corporate management are outside the scope of both GSIM and GSBPM.

38. GSIM is robust, but can readily be adapted and extended to meet users' needs. In order to implement GSIM, you will need to identify the organization-specific information that needs to be integrated into your own extension of GSIM. Examples include preferred platforms and standards, standard documents to be produced when developing a new statistical program, etc.

39. In order to extend GSIM usefully, it is important to use the mechanism provided within GSIM, and to document every extension carefully. The quality of this documentation is fundamental for a successful use of the extensions for communication between all participants in the activities of the organization. Moreover, extensions to GSIM are not for internal use only. They should be submitted to the Modernisation Committee on Standards under the UNECE (support.stat@unece.org), which will keep a record of existing GSIM extensions. Some of them might well be approved as something that should be added to the agreed model.

Implementing GSIM at a technical level
40. In order to effectively implement GSIM at a technical level, it should be implemented at a business level first. GSIM does not provide any standard representation of its own, and is intended to be implemented using existing external standards and models, which support technical implementation. This may involve mapping to internal data models or implementation standards used within an organization.

41. Organizations implementing GSIM will need to map the GSIM objects against their implementation models. In doing this, gaps between GSIM and the implementation standards or internal models will need to be identified. Mappings of GSIM to an organization’s internal models can be shared with the community via the Global Artefact Catalogue (to be developed in 2014).

**GSIM Mapping to SDMX and DDI**

42. For some common standard models, GSIM mappings have been provided. The design of GSIM takes into account the possibility to map to implementation models, such as SDMX or DDI. Such a mapping can be used to establish a link between GSIM and its technical implementation. GSIM thus helps to:

- Compare IT approaches;
- Detect double work in legacy systems
- Avoid double work in new systems.

43. Mapping tables have been developed which show how GSIM objects correspond to their counterparts in the SDMX and DDI standards.

**DDI Profiles for Use with GSIM Implementations**

44. DDI is a very flexible and complex standard, and even within the mappings from GSIM which have been developed, there is still the possibility that two GSIM implementations using DDI might not interoperate. DDI itself provides a mechanism for solving this problem – a feature of the standard known as “DDI Profiles”. It is thought that the publication of standard DDI profiles will facilitate the harmonisation of the use of DDI in statistical organisations.

**What is a DDI Profile?**

45. According to the DDI 3.2 documentation, a DDI profile "describes the subset of valid DDI objects used by an agency for a specified purpose." The profile is documented in an XML format (part of the DDI specification) which allows a set of declarations to be made, identifying specific fields in the DDI which are "Used" or "Not Used". Various other qualifications can be made to restrict or default permitted values for specific elements, and human-readable documentation can be added.

46. Thus, an organization or application can specify exactly how it uses the DDI XML formats. It is anticipated that as organizations implement GSIM, these standard profiles can be used to form a base set of DDI elements for interoperable use.

47. Examples of some of the profiles which have been developed include:

- **BTO Profile**: The first profile is “Basic Technical Objects” (BTO). There are a small set of technical objects that are used in DDI to support identification, referencing, dates, external controlled vocabularies, string content, and the basic features of ISO/IEC 11179-5 (Name, Label, Description). This profile provides guidance for the best practice usage of these technical objects to support interoperability. In particular it identifies those objects which should be considered interoperable and applied in a consistent manner, and those that contain primarily local information.
- **Variable Profile**: This profile covers those DDI objects required to support the Variable, Population and Concept objects in GSIM.
- **Represented Variable Profile**: This profile covers those DDI objects required to support the Represented Variable, Enumerated Value Domain, Described Value Domain and Unit of Measure objects in GSIM.
- **Questionnaire Profile**: This is the definition of a 3.2 DDI profile for business questionnaire - test version. This profile includes all the objects, schemes and 'packaging' information needed to create a complete questionnaire, including Questions, Question Groups and Blocks, Statements, Interviewer Instructions, Controls and Instrument.
- **Codelist Profile**: This profile covers those DDI objects required to support Code Lists and Category Sets in GSIM.

**Common Statistical Production Architecture (CSPA)**

48. CSPA is the industry architecture for the official statistics industry. An industry architecture is a set of agreed common principles and standards designed to promote greater interoperability within and between the different stakeholders that make up an "industry", where an industry is defined as a set of organizations with similar inputs, processes, outputs and goals (in this case official statistics).

49. CSPA builds on and uses existing frameworks, notably the GSBPM and GSIM, as the necessary shared industry vocabulary. CSPA complements and uses these pre-existing frameworks by describing the mechanisms to design, build and share components with well-defined functionality that can be integrated in multiple processes easily.

50. For the purposes of developing sharable, interoperable GSIM-based services and applications, the use of the technical standards SDMX and DDI is envisioned, along with some other standards for specific parts of GSIM such as BPEL (Business Process Execution Language), etc.

51. There is a need to do more than simply refer to relevant existing standards such as SDMX and DDI. The CSPA implementation specification for Statistical Services will specify:

- whether SDMX, DDI or a custom schema should be used for representing a particular GSIM information object, and
- exactly how the chosen schema will be applied for the particular purpose. In many instances there are multiple technically compliant means of achieving the same business purpose, the implementation specification will specify which should be used.
52. Implementation specifications mean CSPA is prescriptive in regard to some practical details. While it would be simpler to align with CSPA if it was less prescriptive, the practical value from alignment would be much less. It is often the case that two developments that have a "common conceptual basis", but were implemented using completely unrelated approaches, are difficult and expensive to make interoperable and/or sharable (if it is possible at all).

53. In addition, an organization that has already implemented a different standard, or a local specification, can "map" their existing approach to the relevant implementation specification – they are not required to "rebuild" from first principles.

54. CSPA implementation specifications specify approaches that will support maximum interoperability/sharability on a cost effective basis. In particular cases it may be difficult for an organization to fully comply with a CSPA implementation specification (due to operational constraints). In these cases, compliance to the extent practical will, generally, still realize significant benefits. In other words, while CSPA implementation specifications set the bar reasonably (but not unreasonably!) high, it is recognized not all implementations may be able to achieve it fully in practice.

1. A fifth group of objects, the "base group", exists mainly for technical modelling purposes, and is generally not shown in communication diagrams such as this one. It is fully described in the specification document.

2. These mapping tables can be found at: http://www1.unece.org/stat/platform/display/metis/Generic+Statistical+Information+Model

3. More information on this on-going work can be found at: http://www1.unece.org/stat/platform/display/metis/Generic+Statistical+Information+Model