

GSIM Communication Paper

Note: To translate this paper into over 50 languages, please see the Automatic translation option at the top of the screen



Word version of document



GSIM Communication Paper v1.0 in Norwegian

(kindly provided by Statistics Norway)

Generic Statistical Information Model (GSIM): Communication Paper for a General Statistical Audience (Version 1.1, December 2013)

About this document

This document provides an overview about the information represented in GSIM, and summaries of how the model will benefit statistical organizations and relationships to other models and standards.

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

Scope

What is GSIM?

Benefits of GSIM for the organization as a whole

GSIM and GSBPM

What does it mean for me?

The Business view

Introduction

1. Across the world statistical organizations undertake similar activities albeit with variations 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). In the past, there was no common way to describe the information that is used. This makes it difficult to communicate clearly within and between statistical organizations and without this there was no foundation for in-depth collaboration, standardization, or the sharing of tools and methods.
2. The Generic Statistical Information Model (GSIM) is the first internationally endorsed reference framework for statistical information. This overarching conceptual framework will play an important part in modernizing, streamlining and aligning the standards and production associated with official statistics at both national and international levels.
3. 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 can be used to guide the development and use of consistent implementation standards or specifications.
4. 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 prepared by the High-Level Group for the Modernization of Statistical Production and Services (HLG), and endorsed by the Conference of European Statisticians¹.
5. The modernization of statistical production is needed in order 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. However, a model alone cannot transform an organization or its processes. In order to meet the future needs of statistical organizations, GSIM is designed to allow for innovative approaches to statistical production to the greatest extent possible. It is one of the main foundations of the Common Statistical Production Architecture², a collaborative initiative to design common and interchangeable services with standard interfaces to support standardisation and modernisation. At the same time, GSIM supports current ways of producing statistics.
6. This paper provides an introduction to GSIM, summarizing the key points for a relatively general statistical audience. For more technical detail, please see the specification document and related material, available on the UNECE web site³.

Scope

7. GSIM provides the information object framework supporting all statistical production processes such as those described in the Generic Statistical Business Process Model (GSBPM)⁴, giving the information objects agreed names, defining them, specifying their essential properties, and indicating their relationships with other information objects. It does not, however, make assumptions about the standards or technologies used to implement the model.
8. GSIM does not include information objects related to 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 GSIM?

9. GSIM 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 (for example, data editing rules). GSIM identifies around 110 information objects, which are grouped into four top-level groups, and are explained in more detail in the specification documentation.

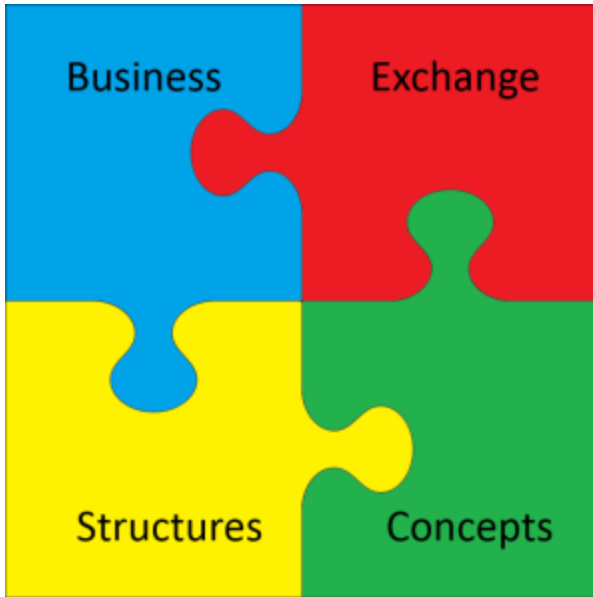


Figure 1. GSIM Top-level information object Groups

10. The four top-level groups are described below:

The **Business** group is used to capture the designs and plans of statistical programs, and the processes undertaken to deliver those programs. This includes the identification of a Statistical Need, the Business Processes that comprise the Statistical Program and the evaluations of them.

The **Exchange** group is used to catalogue the information that comes in and out of a statistical organization via Exchange Channels. It includes objects that describe the collection and dissemination of information.

The **Concepts** group is used to define the meaning of data, providing an understanding of what the data are measuring.

The **Structures** group is used to describe and define the terms used in relation to information and its structure.

11. Figure 2 shows a simplified view of the information objects identified in GSIM. It gives users examples of the objects that are in each of the four top-level groups.

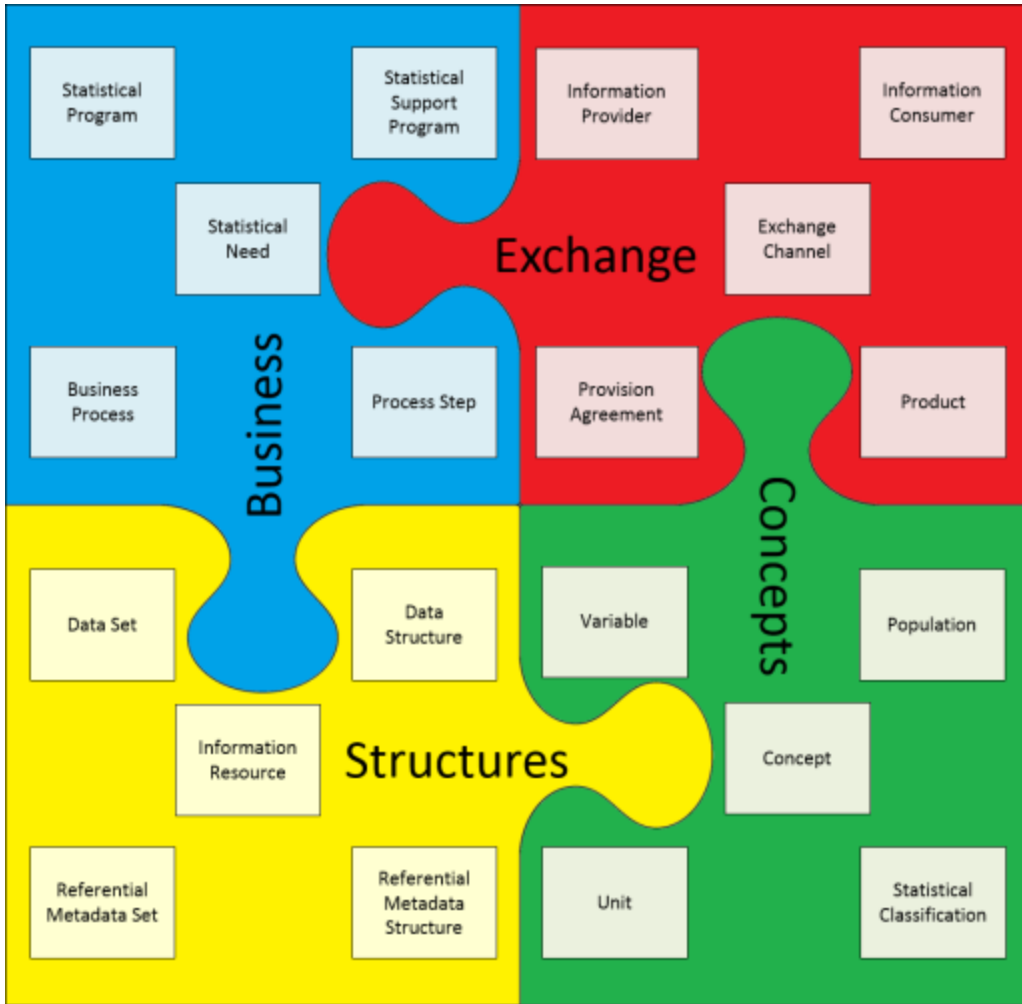


Figure 2. Simplified view of GSIM information objects

12. Figure 3 shows another view of one part of GSIM. This is a slightly more technical view, but still intended to be accessible by a relatively wide audience. Both figures 2 and 3 can be used as a means for communication with users who are interested in examples of the objects and relationships in GSIM.

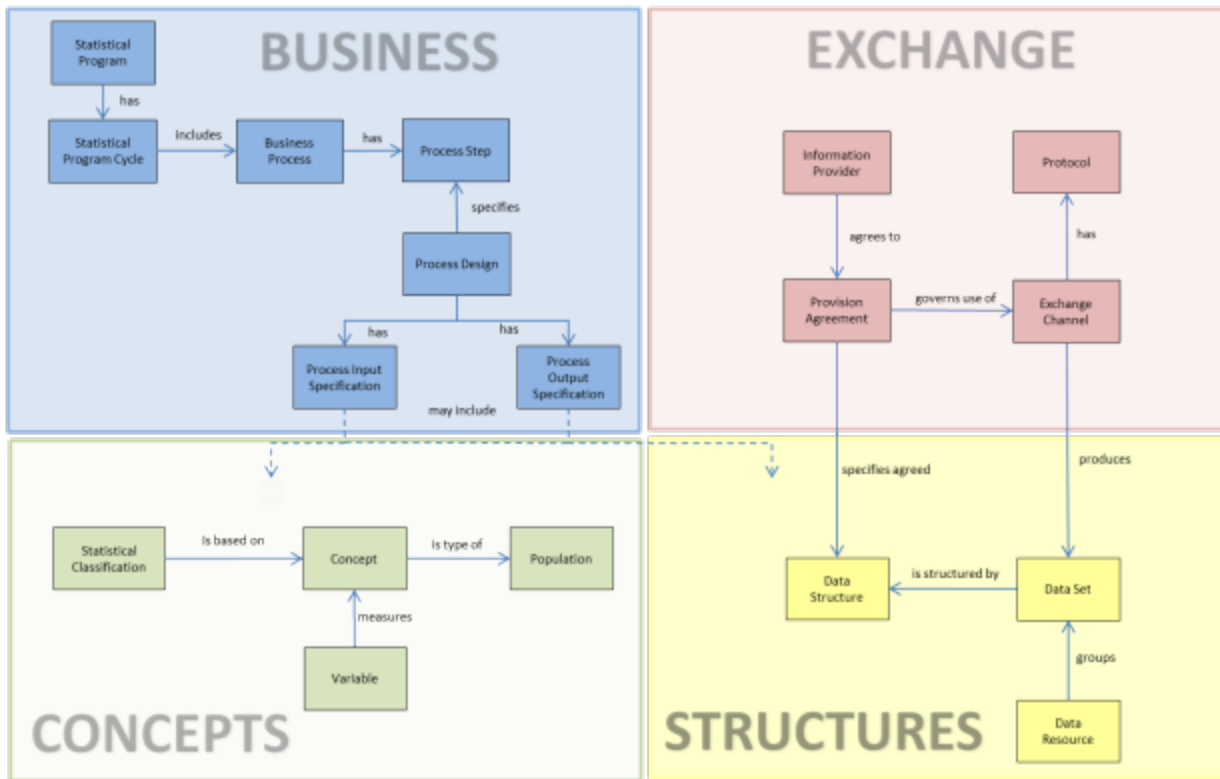


Figure 3. Alternate simplified view of GSIM information objects

13. Figure 3 gives an example of GSIM information objects that tell a story about some of the information that is important in a statistical organisation. Information objects in the GSIM model are given in *italics*.

"A statistical organization initiates a *Statistical Program*. The *Statistical Program* corresponds to an ongoing activity such as a survey or an output series and has a *Statistical Program Cycle* (for example it repeats quarterly or annually).

The *Statistical Program Cycle* will include a set of *Business Processes*. The *Business Processes* consist of a number of *Process Steps* which are specified by a *Process Design*. These *Process Designs* have *Process Input Specifications* and *Process Output Specifications*. The specifications will often be pieces of information that refer to *Concepts* and *Structures* (for example, *Statistical Classification*, *Variable*, *Population*, *Data Structure*, and *Data Set*).

If, for example, the *Business Process* is related to the collection of data, there will be an *Information Provider* who agrees to provide the statistical organisation with data (via a *Provision Agreement*). This *Provision Agreement* specifies an agreed *Data Structure* and governs the *Exchange Channel* used for the incoming information. The *Exchange Channel* could be a *Questionnaire* or an *Administrative Register*. It will receive the information via a particular mechanism (*Protocol*) such as an interview or a data file exchange.

The *Data Set* produced by the *Exchange Channel* will be stored in a *Data Resource* and structured by a *Data Structure*."

Benefits of GSIM for the organization as a whole

14. 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 internally or 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.

Long term benefits

15. GSIM provides a set of standardized information objects, which are the inputs and outputs in the design and production of statistics. By defining objects common to all statistical production, regardless of subject matter, GSIM enables statistical organizations to rethink how their business could be more efficiently organized.

16. 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. For example, some organizations could specialize in seasonal adjustment, time series analysis or data validation, and other organizations could take advantage of this expertise.

17. Implementation of GSIM, in combination with GSBPM, will lead to more important advantages. GSIM could:

- 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;
- Facilitate generation of economies of scale through development of common tools by the community of statistical organizations.

Immediate benefits

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

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

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

20. GSIM can be used by organizations now 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 international or local standards to ensure they meet the broadest needs of the international statistical community.

GSIM and GSBPM

21. 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". As shown in Figure 6, GSIM helps describe GSBPM 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.

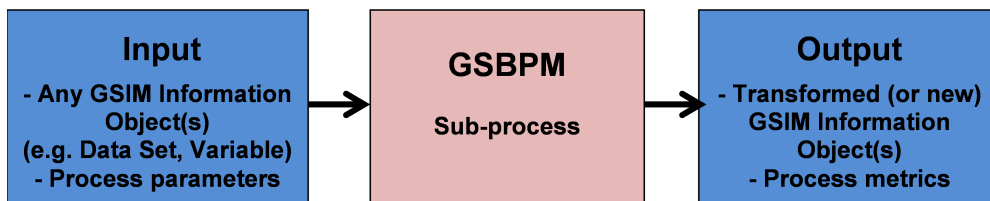


Figure 4. GSIM and GSBPM

22. 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, it is possible (although not ideal) to apply one without the other. In the same way that individual statistical business processes do not use all of the sub-processes described within GSBPM, it is very unlikely that all information objects in the GSIM will be needed in any specific statistical business process.

23. 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 metadata. Metadata management strategies and systems are therefore vital to the operation of GSBPM, and are facilitated by GSIM.

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

- Facilitate the building of efficient metadata driven collection, processing, and dissemination systems;
- Help harmonize statistical computing infrastructures.

25. 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"^[5], that is, that metadata should *uniquely and formally define the content and links between objects and processes in the statistical information system*.

What does it mean for me?

The Business view

26. GSIM will help you to improve your communication with colleagues (both locally and internationally).

27. Communication of subject matter between domains is often poor, making the sharing of concepts, variables, and design components difficult without a complex mapping exercise. GSIM can serve as a common language and will ease communication between:

- Subject matter specialists, methodologists and information technologists;
- Statisticians in different domains of a statistical organization;
- Statisticians in different organizations.

28. GSIM will help you design and understand your processes (and their inputs and outputs) better.

29. For a production cycle, a statistician can design the input and the output, and the process in-between. In GSIM terms, the output and the input can be designed in terms of structures and concepts information objects, and the process in-between can be designed using the business information objects. The structures and concepts objects are provided by subject matter specialists.

30. As seen in Figure 4, if the GSBPM is considered as a frame of reference for statistical production processes, 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 the "Process" phase of the GSBPM). The third level corresponds to a sub-process (for example sub-process 5.3 of the GSBPM – Review and validate). The fourth level consists of the individual building blocks within the sub-process, such as detecting financial values that might be expressed in thousands rather than units.

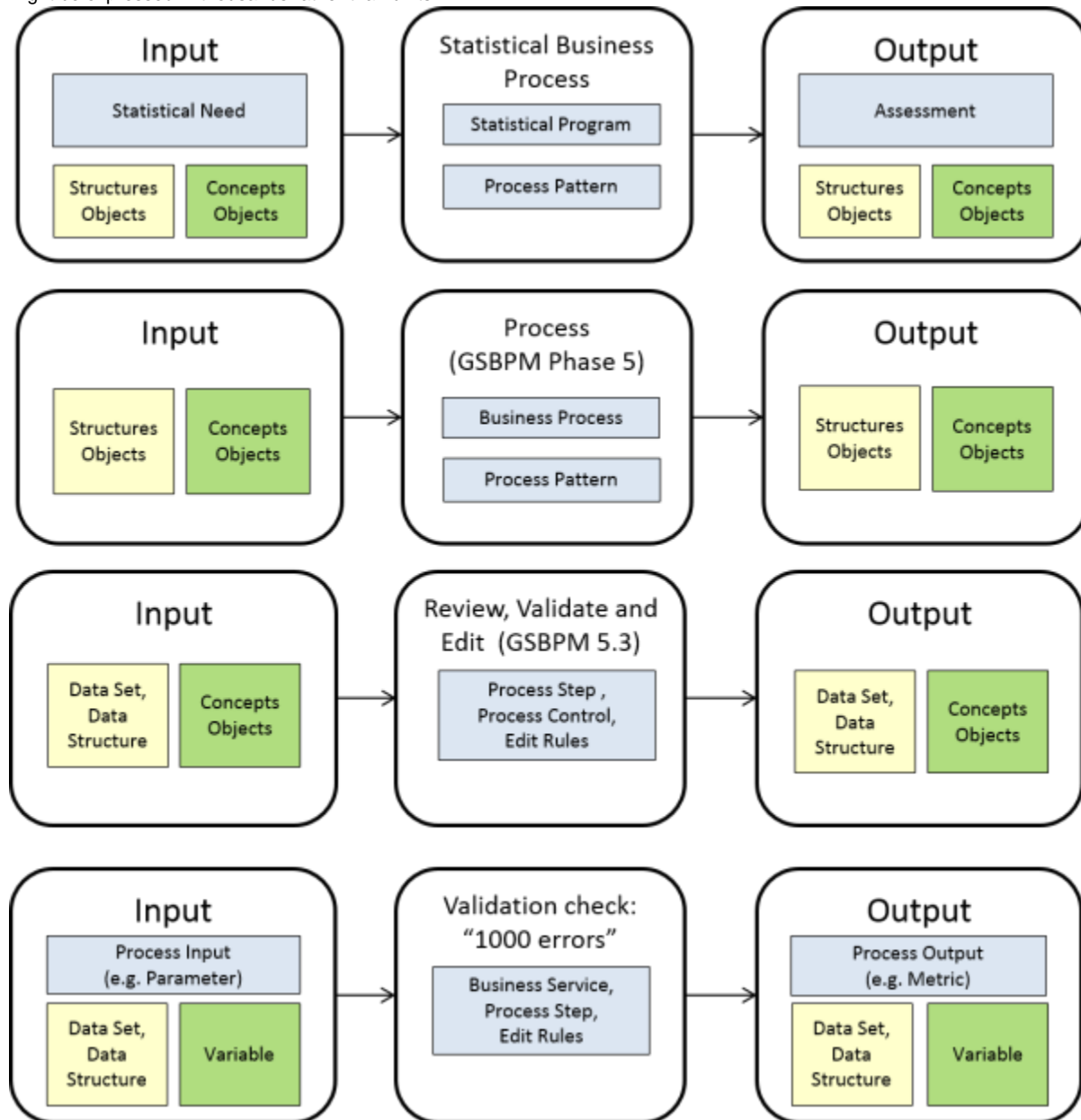


Figure 5. GSIM information objects in context of GSBPM

31. An important issue for statisticians is the problem of single-use design components, which are often recreated or at least modified for each production cycle. GSIM facilitates the description of inputs and outputs at each level of the GSBPM, following the same pattern thus providing a consistent structure to design statistical processes. It supports the design, specification and implementation of harmonized methods and standard technology to create a generalized statistical production system.
32. Using GSIM will enable producing reusable and flexible process building blocks which can be used by statisticians to produce final products of varying complexity, facilitating the production of a wider variety of products and responding more easily to changing client needs.
33. The use of GSIM will reduce workloads as many processes can be repurposed and reused. This means less time spent on repetitive work and more time for innovation.
34. In the long term, GSIM will make statisticians less reliant on information technologists.
35. Statisticians are very much concerned today about the applicability, usability and stability of their methods and technical solutions. In the "stove-pipe" approach to statistical production, subject matter is heavily dependent upon the information technologists in the design, build and production of statistical systems.
36. Statisticians will gain greater control over the design of their processes making them more self-supporting in the design and production of their statistics.
37. Production will be based upon more standardized applications that are more robust to change and less vulnerable to changing personnel. An increase in the use of standardized applications, which can easily be shared across domains, will enable statisticians to more easily work in different domains.

The Information Technology view

38. A main concern for information technologists is the duplication of effort due to the "stove-pipe" organization of statistical production. Unstable and differing requirements from these "stove-pipes" lead to tailor made one-off solutions, whilst a high turnover of IT staff can result in poorly documented and non-standard applications.
39. The introduction of GSIM both at the national and at the international level can already bring short term benefits for information technology specialists. GSIM will provide a common language for information technologists to talk to clients and colleagues both locally and internationally.
40. At the national level, statisticians will become more self-supporting in the design (see Figure 6) and production of their statistics reusing and repurposing harmonized components GSIM will enable more flexible and modular production systems. Production will be based upon more standardized applications that are more robust to change and less vulnerable to changing of IT personnel. An increase in the use of standardized applications, which can easily be shared across domains, will enable the IT specialists to more easily work in different domains.
41. The use of GSIM will reduce the workload as many components can be repurposed and reused. This means less repetitive work and more time for innovation.
42. This will free the IT staff to make more robust applications and explore new ways to better meet the changing needs of the statistical organization and their clients at large. This will include more time for creation of robust, modular, harmonized, well documented processes that comply with the requirements of the Common Statistical Production Architecture.

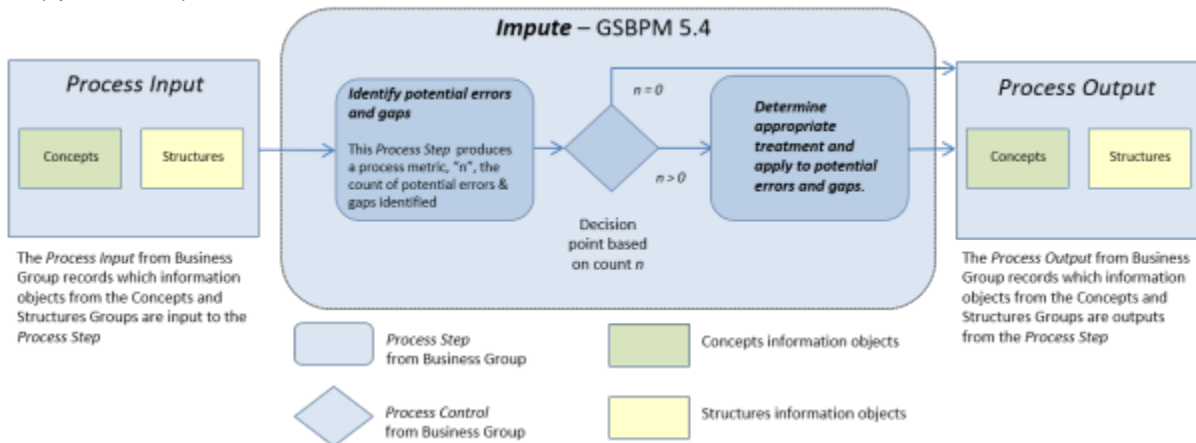


Figure 6. Design your own imputation process

43. At the international level there will be increased possibilities for co-design and co-development of common components based upon more robust user-requirements from a wider user-community. The IT developers will also have access to a larger development community that all speak the same language to describe their statistical information.

SDMX, DDI and other standards

44. As a reference framework of information objects, GSIM has a complementary relationship with standards, such as SDMX (Statistical Data and Metadata eXchange) and DDI (Data Documentation Initiative), which are commonly used to represent and exchange statistical data and metadata.

45. The information objects within GSIM are conceptual; no specific physical representation of the information is prescribed. As a simplified illustration, the name of an organization can be defined as the same concept regardless of whether the information is recorded in a database, in a spreadsheet, in a CSV file, in an XML file or handwritten on a piece of paper.

46. GSIM allows organizations to start with a common language related to the data and metadata used throughout the statistical production process. In this context, GSIM information objects have been mapped to relevant representations in SDMX and DDI.

47. This will help statistical organizations to describe and manage statistical information using a common language while, at a systems level, the information is represented and exchanged in an appropriate and standard technical format.

48. While GSIM information objects can be mapped to SDMX and DDI (and substantial business benefit can be obtained from harnessing these standards), GSIM does not require these standards to be used. Some producers and some users of statistics may decide to use alternative standards for particular purposes. In other cases, producers of statistics may be open to using SDMX and/or DDI but have legacy information systems which are not economical to update for use with these standards.

49. Describing statistical information using GSIM as the common point of reference helps users identify the relationship between two sets of statistical information which are represented differently from a technical perspective.

50. For example, a statistician may receive some data described in DDI and some described in a locally created format. The statistician can relate both of these to GSIM. The statistician will be able to identify which differences are purely technical and which reflect underlying conceptual differences.

51. Once the nature and extent of the differences can be understood, it often proves straightforward to transform the information into a common technical representation (for example, SDMX or DDI) which allows the content to be integrated and explored. This approach ensures that the results of the technical conversion to a common standard are accurately understood, and are sound, from a conceptual perspective.

52. There are a number of synergies between use of GSIM as a reference framework and the application of representation standards such as SDMX and DDI. These synergies have been maximised by design.

53. For example, when determining the set of definitions to be used for information objects within GSIM, existing standards and models were harnessed as key reference sources. While none of these existing sources had the same purpose and scope as GSIM – that is a reference framework of information objects spanning the full statistical production process – the development of each entailed analysing and supporting particular needs and scenarios related to particular types of statistical data and metadata.

54. In this way GSIM benefited from the investment of time in analysis, modelling, testing and refinement when developing these standards and models to their current level of maturity. It also means GSIM does not vary "for no reason" from terms and definitions which are used in existing standards and models. Where it does vary it is for reasons such as existing relevant standards and models being inconsistent internally, with one another and/or statisticians reporting that alternative terms or definitions are more relevant to their business needs. A direct consequence of this was the revision of the Neuchâtel Model for Classifications, to fully align and integrate it with GSIM.

Summary

55. This paper introduces GSIM to people working in statistical organisations. It outlines the benefits of the model as well as how the adoption of the model might impact staff in statistical organisations. The paper also discusses the interaction of GSIM and other frameworks and standards such as GSBPM, DDI and SDMX.

56. For more detailed information on the information objects in GSIM, their definitions, attributes and relations, the GSIM Specification document provides a fine level detail and also discusses the relationship between GSIM and other standards and models. The GSIM wiki page [6](#) also includes links to information about practical implementations, and other resources that might be useful to organisations adopting GSIM as a corporate standard.

[1](#) See: www1.unece.org/stat/platform/display/hlgbas

[2](#) See: <http://www1.unece.org/stat/platform/display/CSPA>

[3](#) See: [http://www1.unece.org/stat/platform/display/metis/Generic+Statistical+Information+Model+\(GSIM\)](http://www1.unece.org/stat/platform/display/metis/Generic+Statistical+Information+Model+(GSIM))

[4](#) See: www.unece.org/stats/gsbpm

[5](#) <http://www1.unece.org/stat/platform/display/metis/The+Common+Metadata+Framework>

[6](#) <http://www1.unece.org/stat/platform/display/metis/Generic+Statistical+Information+Model>