Generic Statistical Information Model (GSIM): Communication Paper for a General Statistical Audience

(Version 2.0, June 2024)

About this document

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

Table of Contents

Introduction

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

2. The Generic Statistical Information Model (GSIM) is the first internationally endorsed reference framework for statistical information. It provides **a set of standardised, 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 organisations. It can provide the foundation for in-depth collaboration, standardisation, or sharing of tools and methods, and thereby, play an important role in modernising, streamlining and aligning standards and production associated with official statistics at both national and international levels.

4. GSIM is one of the cornerstones for modernising 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].

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

6. This paper provides an introduction to GSIM, summarising the key points for a relatively general statistical audience. For more detail, please see the GSIM documents available on the GSIM wiki[2].

Scope

7. GSIM provides the information class framework supporting all statistical business processes such as those described in the Generic Statistical Business Process Model (GSBPM)[3], giving those information classes agreed names, defining them, specifying their essential properties, and indicating their relationships with other information classes. It does not, however, make assumptions about the standards or technologies used to implement the model.

8. GSIM does not include information classes related to activities within an organisation

[1] UNECE Statistics Wikis – HLG-MOS (https://statswiki.unece.org/display/hlgbas)
[2] GitHub (https://unece.github.io/GSIM-2.0/GSIMv2.html) and UNECE Statistics Wikis - GSIM (https://statswiki.unece.org/display/GSIM)
[3] UNECE Statistics Wikis - GSBPM (https://statswiki.unece.org/display/GSBPM)

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).[4]

9. GSIM is a conceptual model and does not prescribe how the information should be implemented. Organisations can choose existing standards (e.g., SDMX, DDI) for the technical implementation (for more, see section "SDMX, DDI and other standards").

What is GSIM?

10. GSIM contains classes which specify information about the real world – "information classes". Examples include data and metadata (such as statistical classifications) as well as the rules and parameter inputs needed for production processes to run (for example, data editing rules). GSIM identifies almost 130 information classes, which are grouped into five top-level groups as in Figure 1.

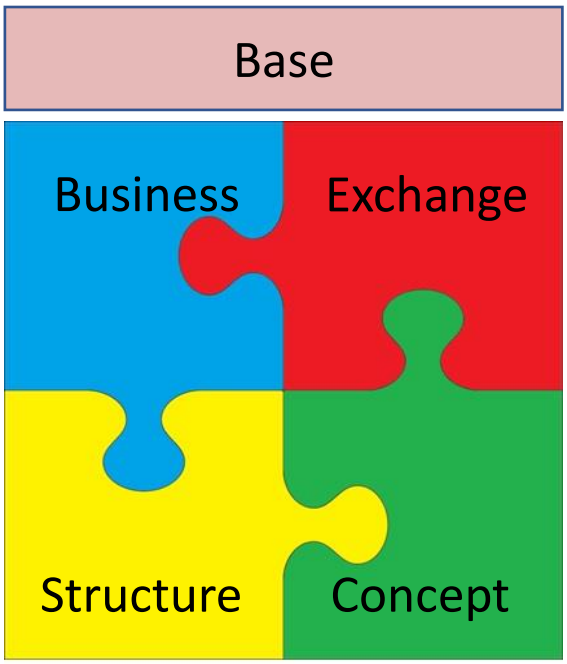

Figure 1. GSIM Top-level information class groups

11. The five top-level groups are described below. Information classes in the GSIM model are given in italics:

- The **Base Group** provides features which are reusable by other information classes to support functionality such as identification and versioning;
- The **Business Group** is used to capture the designs and plans of *Statistical Programmes*, and the processes that are undertaken to deliver those programmes. This includes the identification of a *Statistical Need*, the *Business Processes* that comprise the *Statistical Programme* and the *Assessment* of them;
- The **Exchange Group** is used to catalogue the information that is exchanged within and in/out of a statistical organisation via *Exchange Instruments*. It includes

[4] UNECE Statistics Wikis - GAMSO (https://statswiki.unece.org/display/GAMSO)

information classes that describe the collection/acquisition and dissemination of information;

- The **Concept Group** is used to define the meaning of data, providing an understanding of what the data are measuring;
- The **Structure Group** is used to structure information throughout the statistical business process.

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

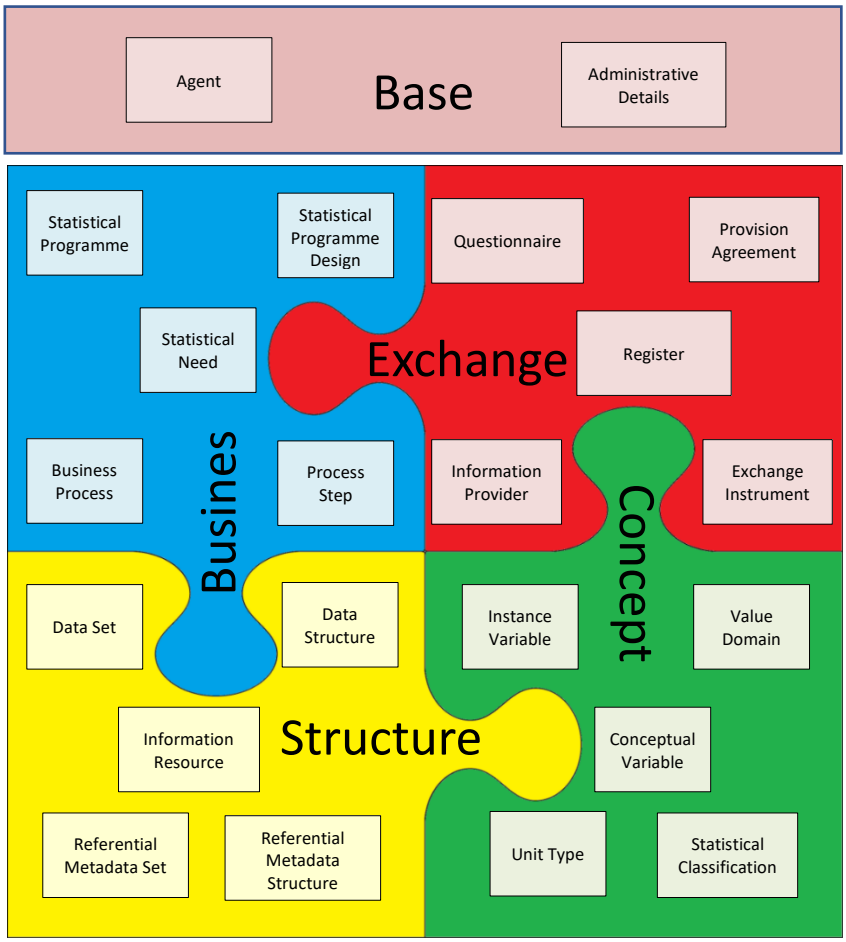

Figure 2. Simplified view of GSIM information classes

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

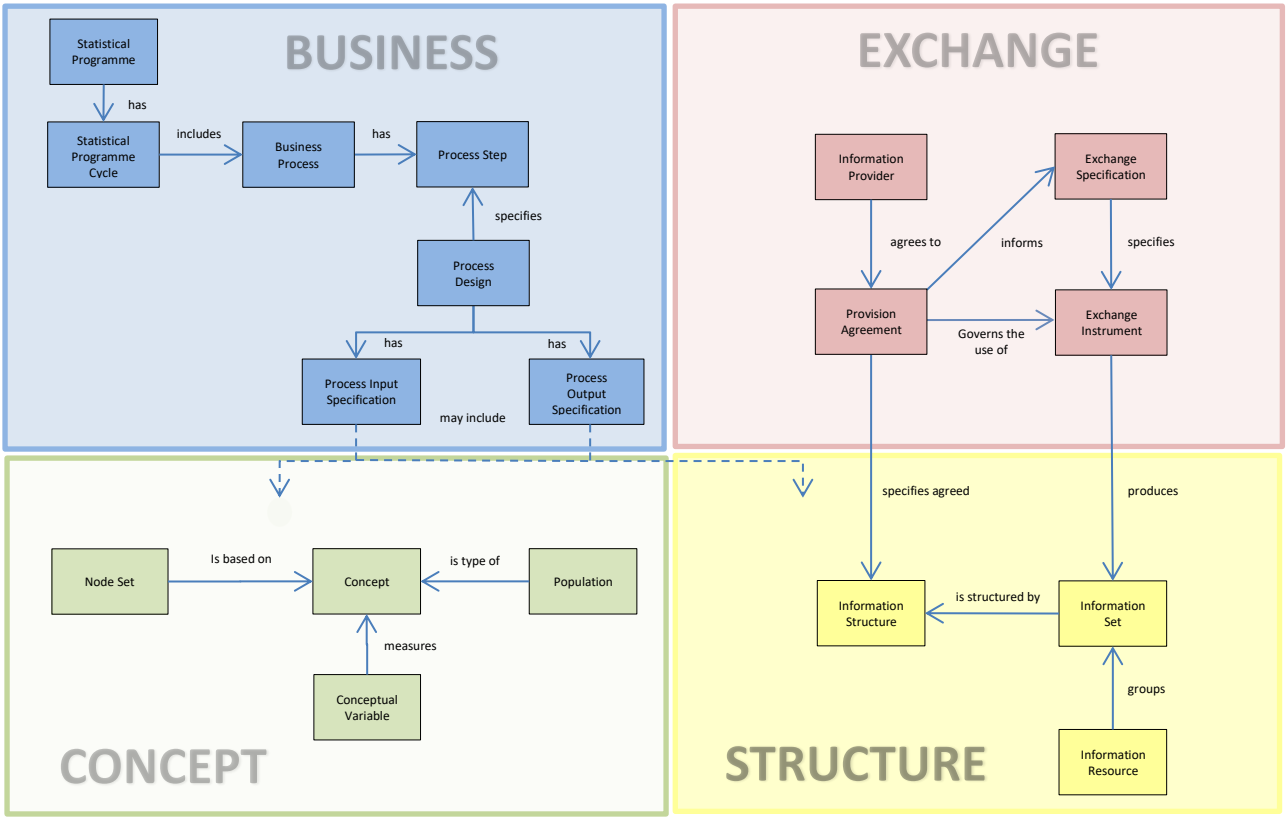

Figure 3. Alternate simplified view of GSIM information classes

14. Figure 3 gives an example of GSIM information classes that tell a story about some of the information that is important in a statistical organisation.

> "A statistical organisation initiates a *Statistical Programme*. The *Statistical Programme* corresponds to an ongoing activity such as a survey or an output series within a *Statistical Programme Cycle* (for example, it repeats quarterly or annually).
>
> The *Statistical Programme Cycle* will include a set of *Business Processes.* The *Business Processes* consist of several *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, *Conceptual Variable*, *Population*, *Information Structure*, or *Information Set*).
>
> If, for example, the *Business Process* is related to the acquisition 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 *Information Structure* (a generalisation of a *Data Structure*) and governs the use of an *Exchange Instrument* which is specified by an *Exchange Specification.*

The *Information Set* (a generalisation of a *Data Set*) produced by the *Exchange Instrument* will be stored in an *Information Resource* and structured by an *Information Structure*."

15. More information about the groups and their information classes can be found in the GSIM wiki and GitHub repository for GSIM 2.0.

Benefits of GSIM for the organisation as a whole

16. It is intended that GSIM may be used by organisations to different degrees. It may be used in some cases only as a model to which organisations refer when communicating internally or with other organisations, to clarify the discussion. In other cases, an organisation 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 organisations that make use of GSIM to its fullest extent may expect to realise the greatest benefits.

Long term benefits

17. The standardised information classes that GSIM provides are the inputs and outputs in the design and production of statistical business processes. These information classes are common to all statistical production, regardless of subject matter, which enables statistical organisations to rethink how their business could be more efficiently organised.

18. 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 specialisation within the international statistical community. For example, some organisations could specialise in seasonal adjustment, time series analysis or data validation, and other organisations could take advantage of this expertise.

19. 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 minimising human intervention in the production process;
- Facilitate the generation of economies of scale through the development of common tools by the community of statistical organisations.

Immediate benefits

20. A significant 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 organisations at national and international levels.

21. Improving communication will result in a more efficient exchange of data and metadata within and between statistical organisations, and with external information providers and consumers.

22. GSIM can be used by organisations to:

- Build capability among staff by using GSIM as a teaching aid that provides a simpler and easier way to understand complex information;
- Validate existing information systems and compare with emerging international best practice and, where appropriate, leverage off international expertise;
- Guide development or updating of international, national or local standards to ensure they meet the broadest needs of the international statistical community.

Relationship with other ModernStats models

23. GSIM has links to several models that have been developed under the auspices of the HLG-MOS to support the modernisation of official statistics (e.g., GSBPM, GAMSO and CSPA, collectively called the "ModernStats" models along with GSIM).

24. GSIM and GSBPM are complementary models for the production and management of statistical information. GSBPM models the statistical business 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 "Edit and impute" and "Calculate aggregates". As shown in Figure 4, GSIM helps describe GSBPM sub-processes by defining the information classes that flow between them, that are created in them, and that are used by them to produce official statistics[5].

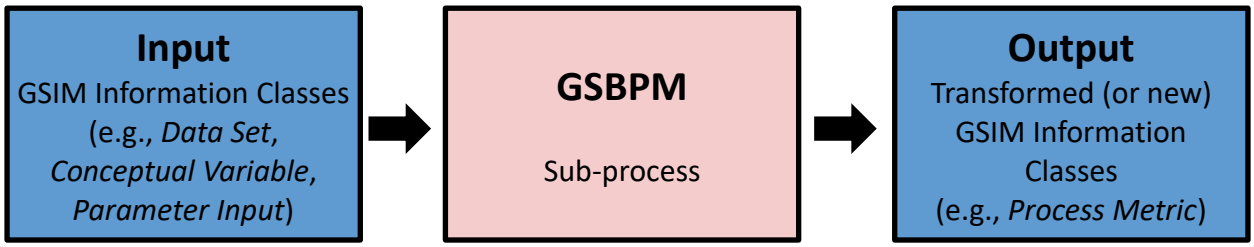

Figure 4. GSIM and GSBPM

25. Greater value will be obtained from GSIM if it is applied in conjunction with GSBPM. 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 the sub-processes described within GSBPM, it is very unlikely that all information classes in the GSIM will be needed in any specific statistical business process.

26. 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 that are developed and

[5] For more information, see https://statswiki.unece.org/display/GSBPM/Information+flow+within+GSBPM+using+GSIM

maintained in GAMSO activity areas are therefore vital to the operation of GSBPM and are facilitated by GSIM.

27. GSIM can also support 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"[6], that is, metadata should uniquely and formally define the content and links between classes and processes in the statistical information system.

28. Describing statistical business processes and their inputs and outputs using the standardised vocabulary of the GSBPM and GSIM supports architecture-management functions by:

- Facilitating the building of efficient metadata-driven collection, processing, and dissemination systems;
- Harmonising statistical computing infrastructures;
- Designing standardised methods and functions for IT applications/tools supporting statistical business processes.

29. In order to meet the future needs of statistical organisations, GSIM is designed to allow for innovative approaches to statistical production to the greatest extent possible. The Common Statistical Production Architecture (CSPA)[7] is an example of a collaborative initiative to design common and interchangeable services with standard interfaces that support standardisation and modernisation. GSIM is one of the main foundations of CSPA, which uses GSIM as a common reference when defining the information input into, output from and support of business processes. Using GSIM as a common language increases the ability to compare information within and between statistical organisations and hence facilitate the development of harmonised and re-usable services and components. At the same time, GSIM also supports current ways of producing statistics.

What does it mean for me?

The business view

30. GSIM will help you in **effectively managing your organisation's business architecture** by providing a standardised list of information classes used in statistical production. Mapping actual inputs and outputs of statistical production onto information classes of the GSIM supports standardisation across subject matter domains.

31. GSIM will help you to improve your communication with colleagues (both locally and internationally). 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, architects, information technologists, quality managers and metadata managers;

[6] UNECE Statistics Wikis – Common Metadata Framework (https://statswiki.unece.org/display/hlgbas/The+Common+Metadata+Framework)
[7] UNECE Statistics Wikis – CSPA (https://statswiki.unece.org/display/CSPA)

- Statisticians in different organisations, or different domains of a statistical organisation.

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

33. 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 classes, and the process in-between can be designed using the business information classes. The structures and concepts classes are provided by subject matter specialists.

34. Figure 5 illustrates the relevance of GSIM classes to such inputs, processes and outputs, in different statistical business process scenarios of successively greater granularity. In that figure the GSBPM is considered as a frame of reference for statistical production processes, where:

- The scenario at the top of that figure can be considered as equivalent to the statistical production process as a whole.
- The next level down corresponds to a single 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 Tasks that are the building blocks within the sub-process, such as Task 5.3.2 Select validation methods based on design[8].

35. 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 thereby supports the design, specification and implementation of harmonised methods and standard technology to create a generalised statistical production system.

36. Using GSIM will enable creating **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.

37. The use of GSIM, in combination with other ModernStats models, will reduce workloads as many processes can be repurposed and reused. This means less time spent on repetitive work and more time for innovation.

38. In the long term, GSIM, in combination with other ModernStats models, will make statisticians less reliant on information technologists, as tools can be designed and developed to be parameterised for dealing with projects from different domains.

39. 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

[8] For further details about GSBPM tasks (https://statswiki.unece.org/display/GSBPM/GSBPM+Tasks)

production, the subject matter specialist is heavily dependent upon the information technologists in the design, build and production of statistical systems.

40. Statisticians will gain greater control over the design of their processes, making them more self-supporting in the design and production of their statistics.

41. Production will be based upon more standardised applications that are more robust to change and less vulnerable to changing personnel. An increase in the use of standardised applications implementing standardised methods, which can easily be shared across domains, will enable statisticians to more easily work in different domains.

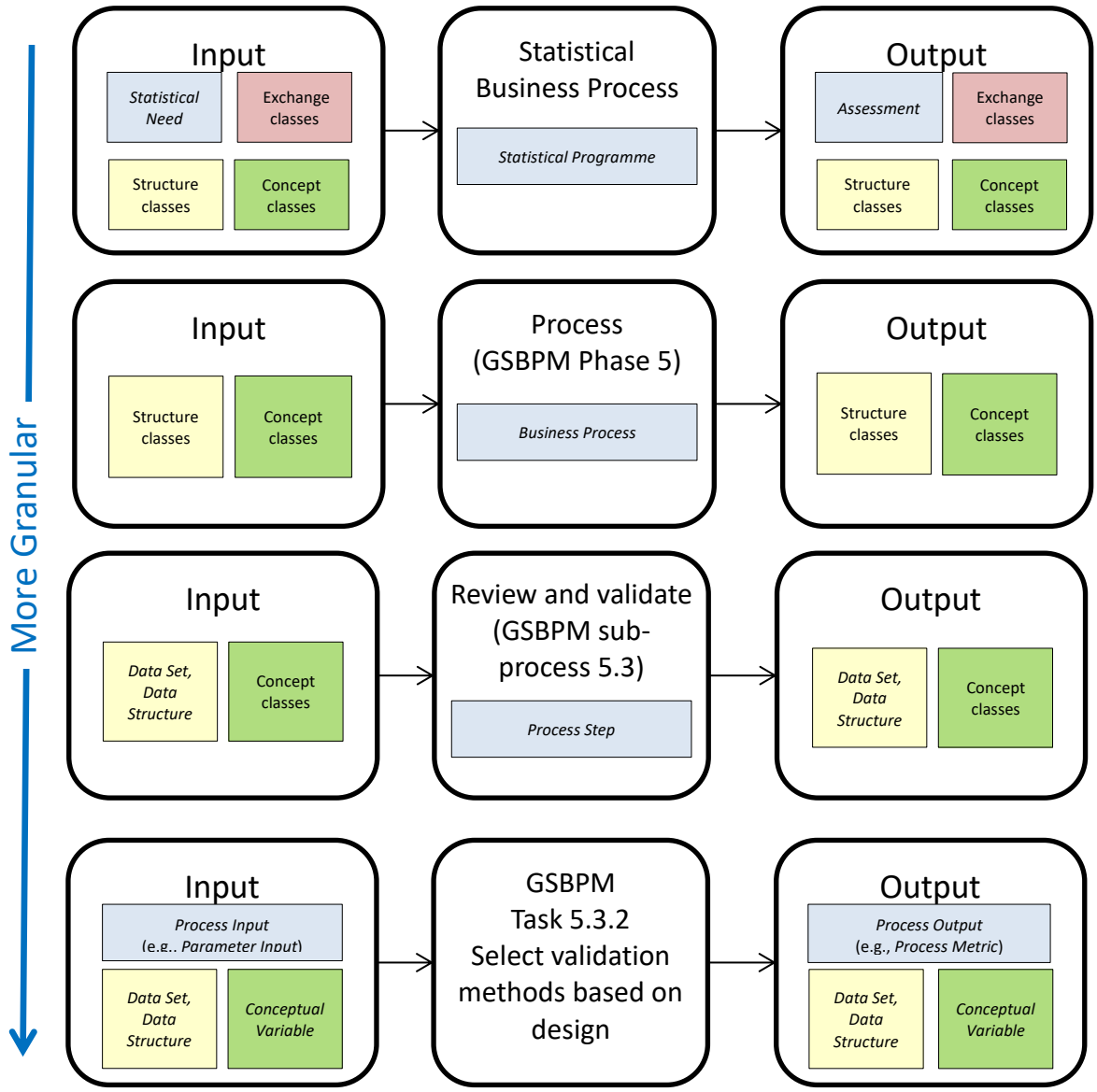

Figure 5. Illustrative mapping[9] of GSIM information classes to inputs, processes and outputs in different statistical business process scenarios (with reference to GSBPM)

[9] Note that this mapping is for illustrative purpose only. User can choose to map GSIM classes differently depending on the use case and intended level of granularity (e.g., *Process Step* could be mapped to an entire GSBPM phase).

The information technology view

42. The main concern for information technologists is the duplication of effort due to the "stove-pipe" organisation of statistical production. Different requirements from these "stove-pipes" lead to tailor-made one-off solutions, whilst turnover of Information Technology (IT) staff can result in poorly documented and non-standard applications.

43. The introduction of GSIM both at the national and at the international level can bring fairly immediate benefits for IT specialists, as GSIM will provide a common language for information technologists to talk to clients and colleagues locally, nationally and internationally. The semantic network of information classes provided by GSIM helps to understand their intrinsic relations and to improve the design of platforms and systems that are more interoperable with each other.

44. At the national level, statisticians will become more self-supporting in the design (see Figure 6) and the production of their statistics by reusing and repurposing harmonised components, as GSIM, together with other ModernStats models, will **enable more flexible and modular production systems**. Production will be based upon more standardised applications that are more robust to change and less vulnerable to changing of IT personnel. An increase in the use of such standardised applications, which can easily be shared across domains, will enable the IT specialists to more easily work in different domains.

45. As mentioned in the preceding *business view* discussion, 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.

46. This will free the IT staff to make more robust applications and explore new ways to better meet the changing needs of the statistical organisation and their clients at large. This includes more time for the creation of robust, modular, harmonised, well-documented processes components, that could be based on the requirements of CSPA.

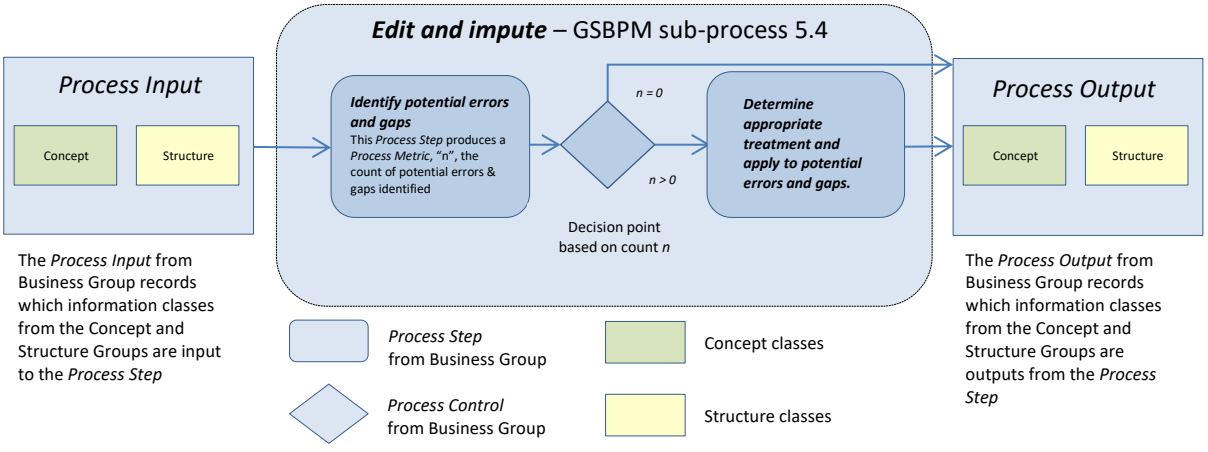

Figure 6. Design your own imputation process

47. 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.

The management view

48. Using GSIM (especially when it is used with GSBPM together) as a base for standardised information classes can support various management activities covered by GAMSO activity areas:

- For the top management - making decisions and planning on statistical programmes, and their controlling activities;
- For financial management – controlling systems for statistical business processes and calculating costs of statistical products;
- For quality management – designing a quality indicator, implementing a quality framework and monitoring product and process quality;
- For methodology management – designing, standardising and maintaining methodologies;
- For information management – managing metadata and data management system/tools and compiling a data management strategy.

SDMX, DDI and other standards

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

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

51. GSIM allows organisations to start with a common language related to the data and metadata used throughout the statistical business process. This will help statistical organisations 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.

52. While GSIM information classes can be mapped to SDMX and DDI[10] (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 those standards.

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

[10] UNECE Statistics Wikis – GSIM and standards, for examples (https://statswiki.unece.org/display/gsim/GSIM+and+standards)

54. 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, and will then be able to identify which differences are purely technical and which reflect underlying conceptual differences.

55. Once the nature and extent of these 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.

56. There are a number of synergies between the 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.

57. For example, when determining the set of definitions to be used for information classes 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 classes spanning the full statistical business process – the development of each entailed analysing and supporting particular needs and scenarios related to particular types of statistical data and metadata.

58. 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 needlessly 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.

Summary and concluding remarks

59. This paper introduces GSIM to those who work in statistical organisations. It outlines the benefits of the model as well as how the adoption of the model might benefit staff in those organisations. The paper also discusses the interaction of GSIM and other frameworks and models such as GSBPM, GAMSO, CSPA, DDI and SDMX.

60. In addition to providing a reference framework for statistical information, GSIM aims to provide ideas and help for modernisation of official statistics. A substantial amount of knowledge and experience from various statistical organisations sits behind its development over the years. With more and more statistical organisations using or planning to use GSIM as a part of their modernisation programme, the user base of the model is steadily growing. Exchange of knowledge and lessons learned benefits new users and experienced users alike and feedback from users is integral for the revisions of the model. GSIM was created by the official statistics community for the official statistics community, all GSIM users are invited to share experiences and thoughts on the GSIM GitHub repository[11] and collectively shape the future development of the model.

[11] GSIM 2.0 GitHub repository (https://unece.github.io/GSIM-2.0/GSIMv2.html)