United Nations ECE/CES/2014/2

Economic and Social Council

Distr.: General
28 January 2014

Original: English

Economic Commission for Europe

Conference of European Statisticians

Sixty-second plenary session
Paris, 9-11 April 2014
Item 5 of the provisional agenda
Work of the High-level Group for the Modernization of Statistical Production and Services

Generic Statistical Information Model (GSIM): Communication paper for a general statistical audience

Prepared by the High-Level Group for the Modernization of Statistical Production and Services

Summary

This paper introduces the Generic Statistical Information Model 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 the Generic Statistical Information Model and other frameworks and standards such as the Generic Statistical Business Process Model, the Data Documentation Initiative and the Statistical Data and Metadata Exchange standard. More information on this information model is available in the Generic Statistical Information Model Specification document, excerpts provided in the Annex, and on the related wiki page[1] that might be useful to organisations adopting this information model as a corporate standard.

The document is submitted to the Conference of European Statisticians for information.

[1] http://www1.unece.org/stat/platform/display/metis/Generic+Statistical+Information+Model

Contents

I. 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 Statisticians[2].

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 Architecture[3], 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 site[4].

[2] See: www1.unece.org/stat/platform/display/hlgbas
[3] See: http://www1.unece.org/stat/platform/display/CSPA
[4] See: http://www1.unece.org/stat/platform/display/metis/Generic+Statistical+Information+Model+(GSIM)

II. Scope

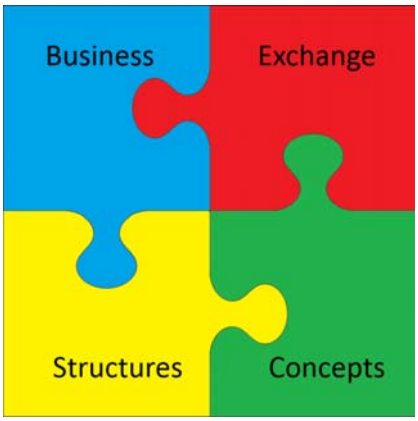
7. GSIM provides the information object framework supporting all statistical production processes such as those described in the Generic Statistical Business Process Model (GSBPM)[5], giving the information objects agreed names, defining them, specifying their essential properties, and indicating their 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.

III. What is Generic Statistical Information Model?

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:

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

(b) 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.

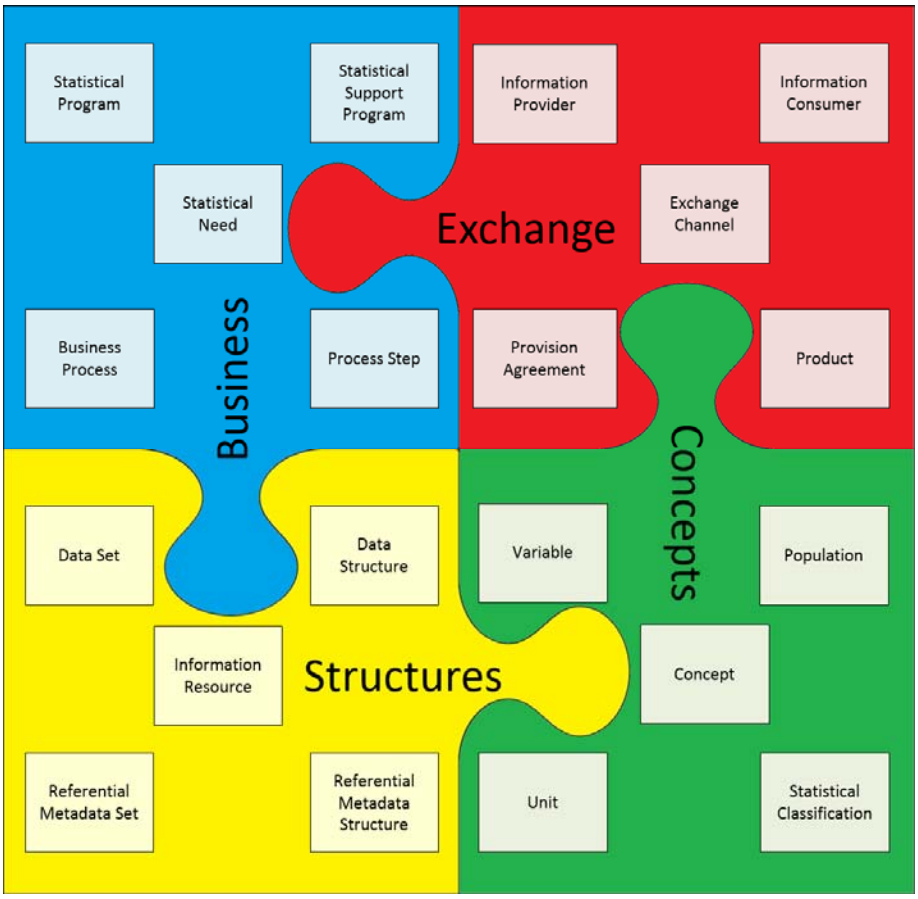
(c) The *concepts group* is used to define the meaning of data, providing an understanding of what the data are measuring.

(d) The *structures group* is used to describe and define the terms used in relation to information and its structure.

[5] See: www.unece.org/stats/gsbpm, document number (ECE/CES/2014/1).

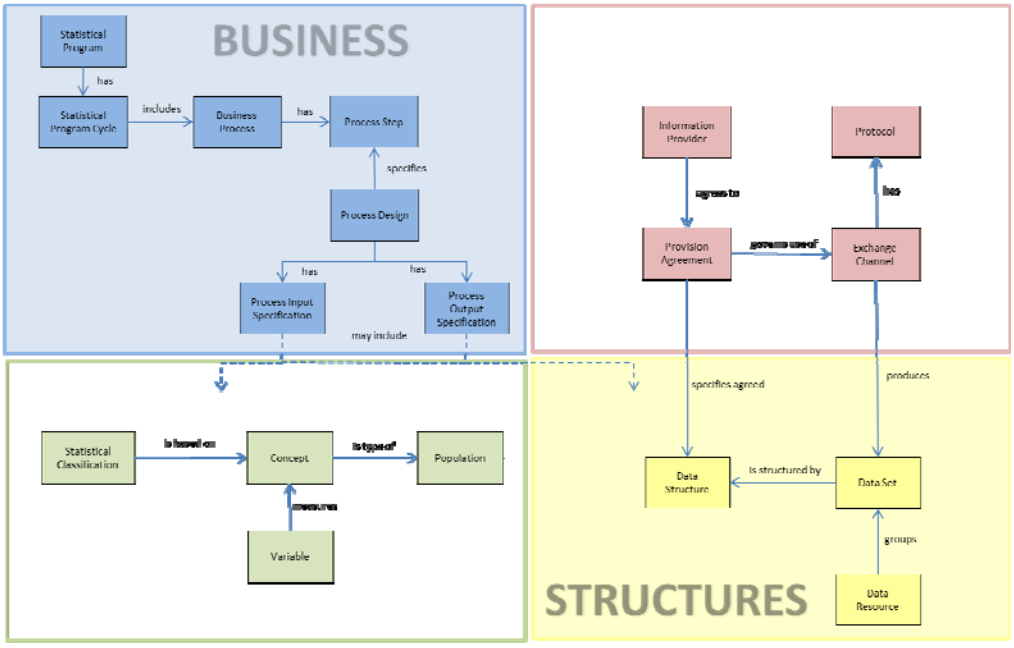
11. Figure 2 shows a simplified view of the information objects identified in GSIM. It gives users examples of the objects that are in each of the four top-level groups.

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 Programme*. The *Statistical Programme* corresponds to an on-going activity such as a survey or an output series and has a *Statistical Program Cycle* (for example it repeats quarterly or annually).
>
> The Statistical Programme 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.*"

IV. Benefits of the Generic Statistical Information Model 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.

A. 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.

B. 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.

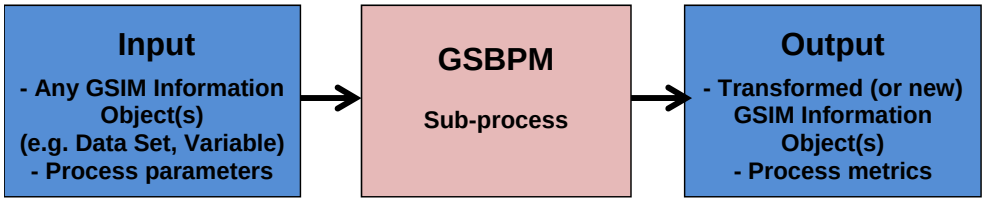
20. GSIM can be used by organizations now to:

- 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.

V. Generic Statistical Information Model and Generic Statistical Business Process Model

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.
The relation of 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"[6], that is, that metadata should *uniquely and formally define the content and links between objects and processes in the statistical information system.*

[6] http://www1.unece.org/stat/platform/display/metis/The+Common+Metadata+Framework

VI. What does it mean for me?

A. 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:

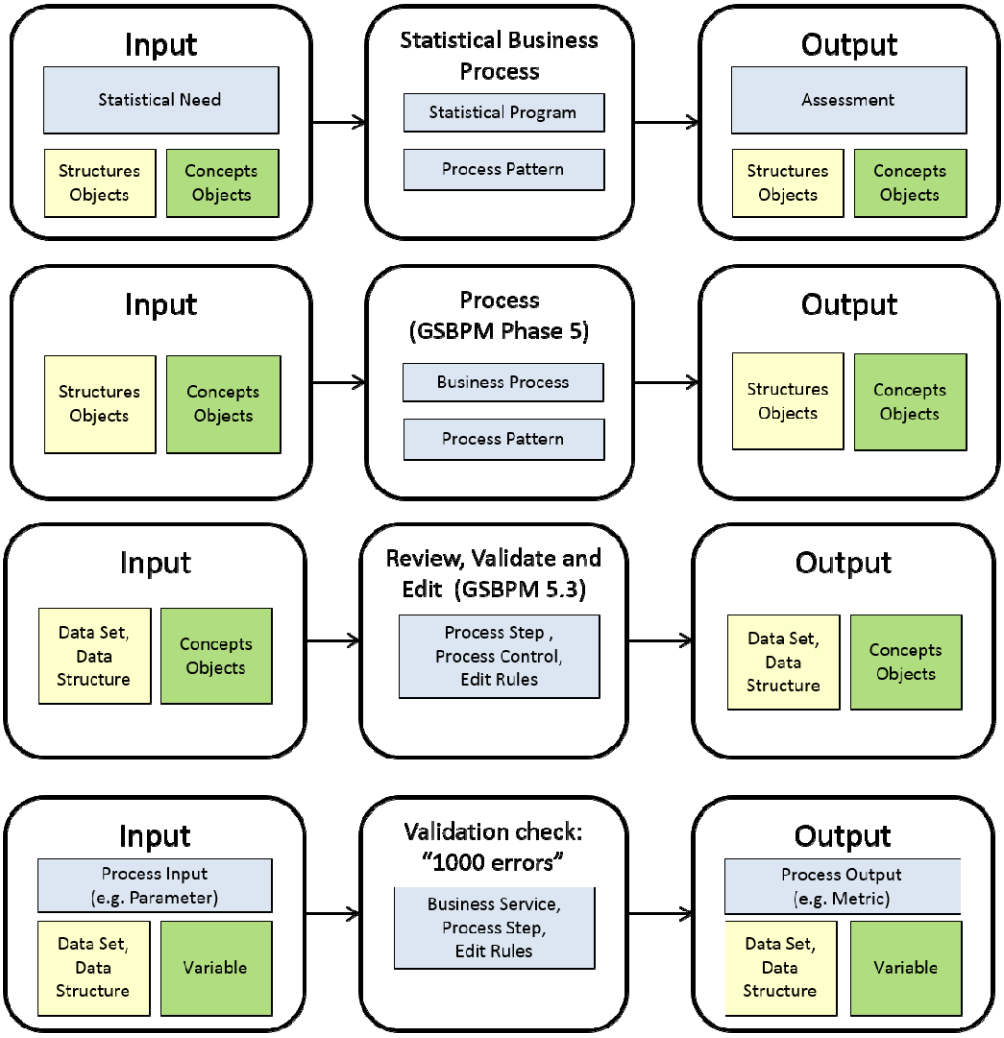
- Subject matter specialists, methodologists and information technologists;
- Statisticians in different domains of a statistical organization;
- Statisticians in different organizations.

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.

B. 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.

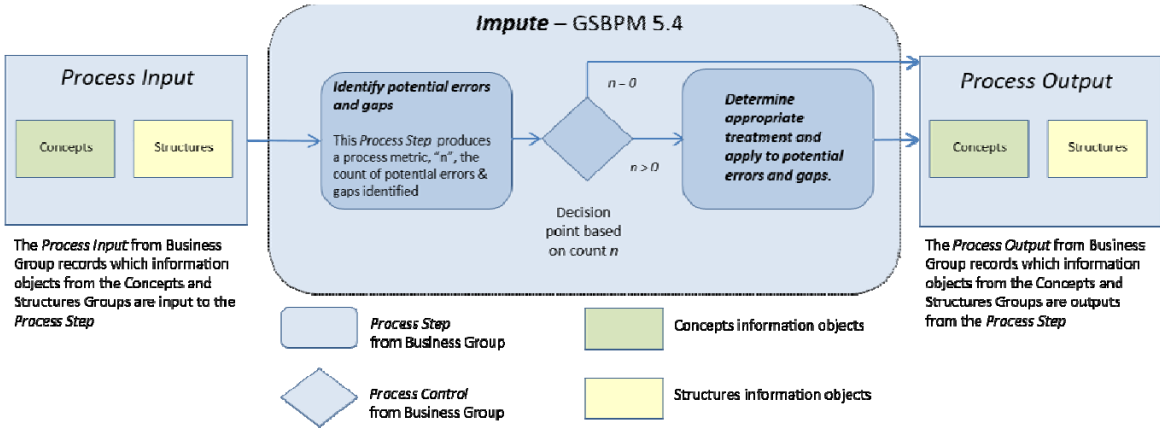
39. The introduction of GSIM both at the national and at the international level can already bring short term benefits for information technology specialists. GSIM will provide a common language for information technologists to talk to clients and colleagues both locally and internationally.

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.

VII. Statistical Data and Metadata Exchange standard, Data Documentation Initiative and other standards

44. As a reference framework of information objects, 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.

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 spread sheet, in a Comma Separated Value(s) (CSV) file, in an Extended mark-up language (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 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.

Annex

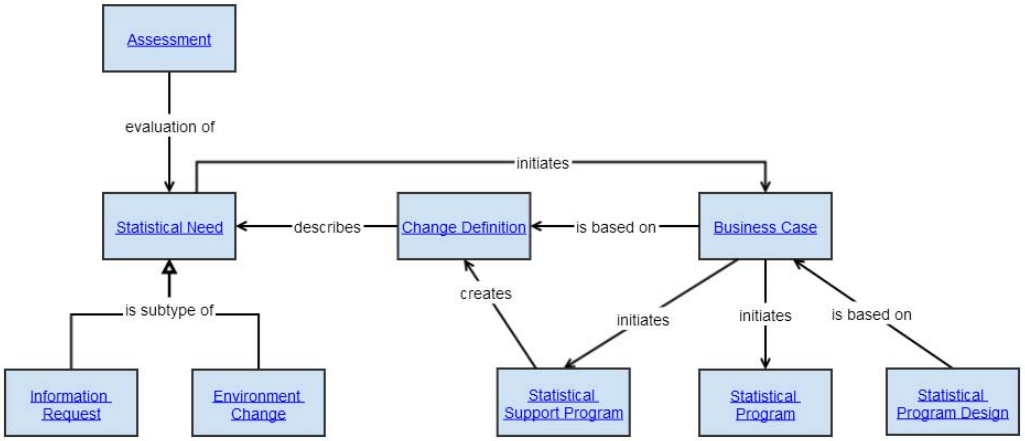
Excerpts from Generic Statistical Information Model Specification document[7]

A. Information in the Statistical Business Process

1. This section looks at different ways that information objects are used within the statistical business process. It considers eight different scenarios, identifying the information objects used and the relationships between those objects.

1. Identifying and Evaluating Statistical Needs

Figure 1.
Identify and Evaluate Statistical Needs



2. An organization will react and change due to a variety of needs. A Statistical Need presents itself to the statistical organization in the form of an Environment Change or an Information Request.

3. Environment Change indicates that there needs to be an externally motivated change. This may be specific to the organization in the form of reduced budget or new demands from stakeholders, or may be a broader change such as the availability of new methodology or technology.

4. When an organization receives an Information Request this will identify the information that a person or organization in the user community requires for a particular purpose. This community may include users within the organization as well as external to it. For example, a the team responsible for compiling National Accounts may need a new Business Process to be initiated to produce new inputs to their compilation process. This request will commonly be defined in terms of a Subject Field that defines what the user wants to measure. When an Information Request is received it will be discussed and clarified with the user. Once clarified, a search will be done to check if the data already

[7] Note: This appendix is not the full GSIM Specification text, it only includes excerpts of that document.

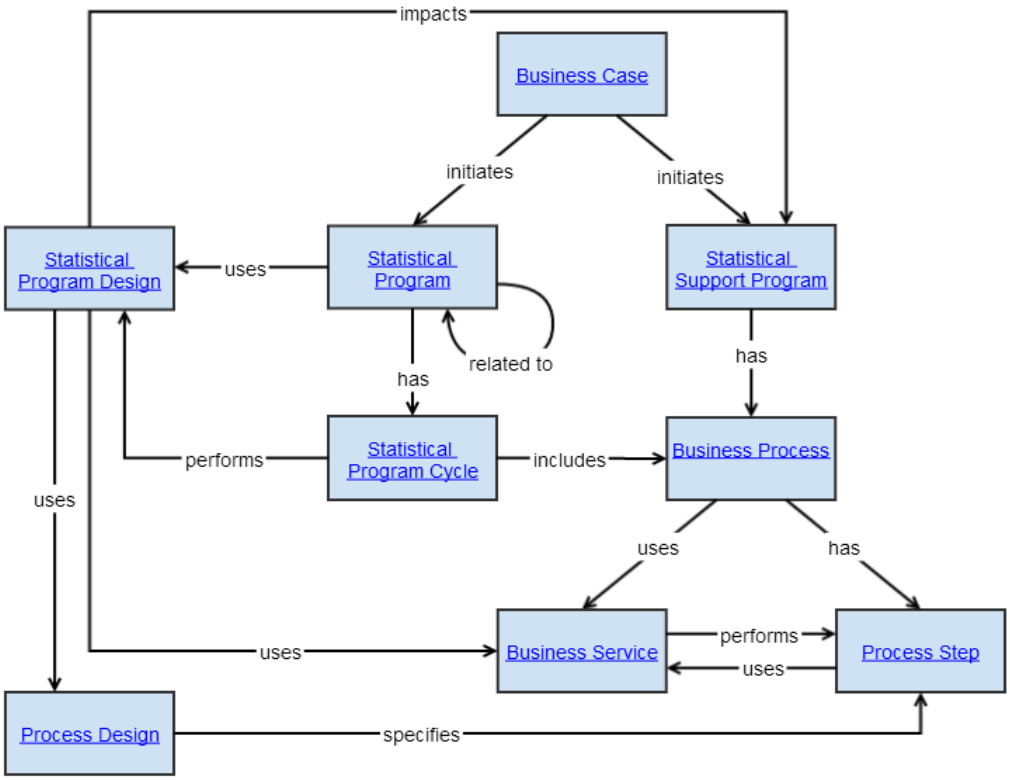
exist. Discovering these Data Sets may be enabled by searching for Concepts and Classifications. Each of these activities are described by a Process Step.

5. The Statistical Need - whether an Information Request or Environmental Change - will be formalized into a Change Definition, typically created by a Statistical Support Programme (a "statistical change programme"). The Change Definition identifies the specific nature of the change in terms of its impacts on the organization or specific Statistical Programmes or Statistical Support Programmes. This Change Definition is used as an input into a Business Case. A successful outcome will either initiate a new Statistical Programme or a new Statistical Support Programme that will create a new Statistical Programme Design that redefines the way an existing Statistical Programme is carried out.

6. A Statistical Need can also be internally driven. At any point in the statistical business process, an organization may undertake an evaluation to determine utility or effectiveness of the business process or its inputs and outputs. An Assessment will be undertaken to evaluate any resources, processes or outputs and may refer to any object described in the model. Assessments include gap analyses undertaken in the context of Business Cases and evaluations undertaken to determine whether a statistical output meets the need for which it was first created.

2. Designing and Managing Statistical Programmes

Figure 2.
Design and Manage Statistical Programme



7. A statistical organization will respond to a perceived Statistical Need by creating a Business Case. Responding to the Business Case will involve one of three things: the creation of a new Statistical Support Programme, the creation of a new Statistical Programme, or the evolution of an existing Statistical Program Design to be implemented by an existing Statistical Programme.

8. Statistical Support Programmes undertake the activities of the statistical organization such as statistical change programmes, data management programmes, metadata management programmes, methodological research programmes, etc. A good example is a programme which manages classifications.

9. Statistical Programmes are those programmes that an organization undertakes to produce statistics (for example, a retail trade survey). Statistical Programmes are cyclical - they perform cycles of collection, production and dissemination of products. Each such cycle is represented by a Statistical Programme Cycle object. The Statistical Programme Cycle is a repeating activity to produce statistics at a particular point in time (for example, the retail trade survey for March 2012).

10. Statistical Programmes require Statistical Programme Designs to achieve their objectives. These designs cover the design of all activities to be undertaken, notably at the level of Business Processes. Within a Statistical Programme Cycle, several Business Processes would typically be performed. These can be understood to correspond to the processes and sub-processes found in the Generic Statistical Business Process Model (GSBPM). These Business Processes may be repeated within a cycle. Each iteration can be made up of multiple activities of the same or different types. As an example of this, within a single cycle, the Statistical Programme might perform three iterations of data collection and processing, then analyse the data and disseminate the resulting statistical Products. Each of these activities could be understood to be a separate Business Process.

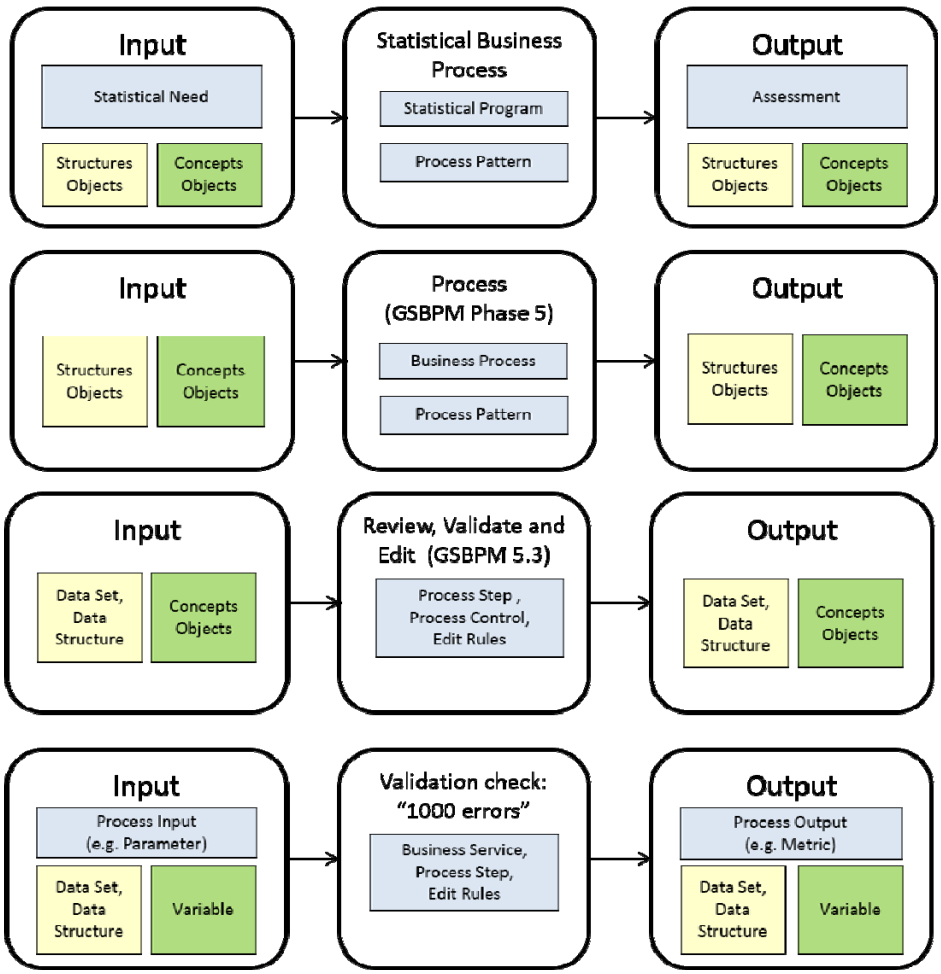
11. The Statistical Programme Design specifies the way in which Business Processes will be conducted. This includes the use of re-usable Business Services (possibly sourced from outside the statistical organization), or through the design and use of more traditional processes. In the latter case, Process Design objects would be used to specify Process Steps. (Although re-usable Business Services are also specified by Process Designs and Process Steps, these will already exist, and not need new design work as part of the Statistical Programme Design.)

12. It should be noted that Statistical Programme Designs specify what Process Steps will need Process Designs, and also which Business Services would be used, but do not do the low-level specification of how such Process Steps and Business Services are executed. These specifications are found in the Process Design object.

3. Designing Process Steps

13. Before explaining the objects which GSIM uses to represent the design of Process Steps, it is important to discuss the nature of processes more generally. The types of objects provided by GSIM perform specific functions. In GSIM, Business Processes have Process Steps. Each Process Step can be as "large scale" or "small scale" as the designer of a particular Business Process chooses (see Figure below).

Figure 3.
Process Steps can be as large or small as needed

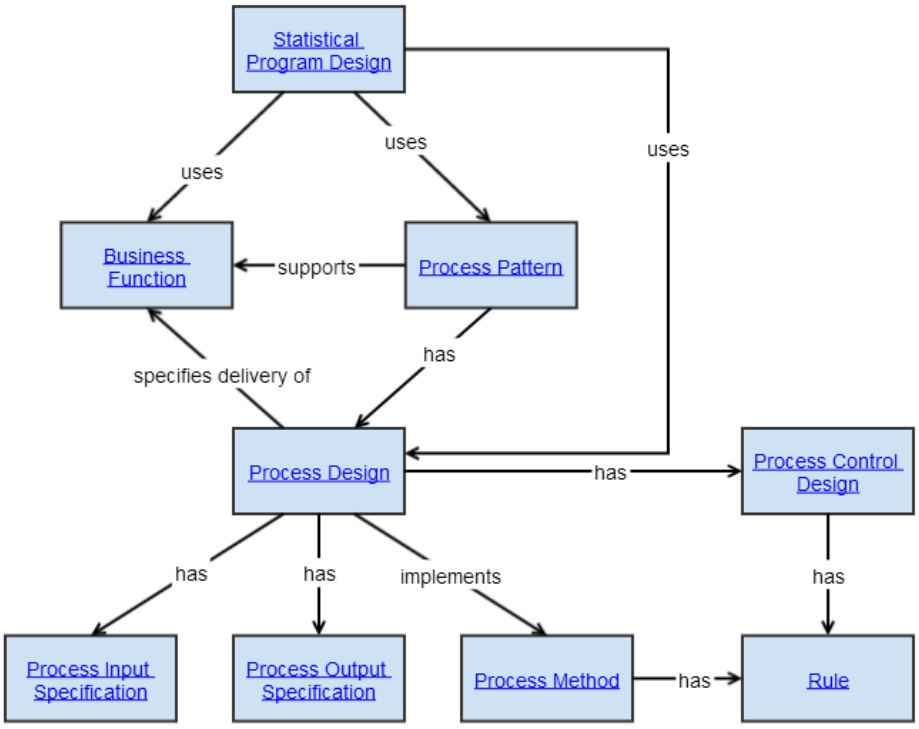


14. Process Steps can contain "sub-steps", those "sub-steps" can contain further "sub-steps" within them and so on indefinitely. Typically, the outputs of one Process Step become inputs to the next Process Step. There can also be conditional flow logic applied to the sequence of Process Steps, based on parameters which have been passed in, or conditions met by the outputs of a previous Process Step.

15. The design of a Process Step thus can be understood to use other Process Steps and even other Business Services which have already been designed and made available for re-use. In a more traditional scenario, the Process Step is designed and then executed. In future, it is foreseen that re-usable Business Services will be increasingly common, having been designed and implemented by another external organization. The next sections describe these two scenarios.

A. Designing Process Steps

Figure 4.
Design Process Steps



16. A Statistical Programme Design is associated with a top level Process Step whose Process Design contains all the sub-steps and process flows required to put that Statistical Programme into effect. Each Process Step in a statistical Business Process has been included to serve some purpose. This is captured as the Business Function associated with the Process Step. An example of a Business Function could be "impute missing values in the data". In order to support this Business Function, an imputation process is needed, which will require a Process Design.

17. In line with the GSIM design principle of separating design and production, GSIM assumes that Process Steps will be designed during a design phase. Having divided a planned statistical Business Process into Process Steps, the next requirement is to specify a Process Design for each step. The Process Design identifies how each Process Step will be performed. A Process Design may use a Process Pattern which is a nominated set of Process Designs and associated flows (Process Control Designs) which have been highlighted for reuse.

18. Process Designs specify several things: they identify the different types of inputs and outputs represented by the Process Input Specification and Process Output Specification. Examples of Process Inputs include data, metadata such as Statistical Classifications, imputation and editing Rules, parameters, etc. Process Outputs can be reports of various types (processing metrics, reports about data validation and quality, etc.), edited Data Sets, new Data Sets, new or revised instances of metadata, etc.

19. To continue the example, the process designer would specify the inputs in the Process Input Specification as imputation Rules and the Data Set for which imputation is desired. The Process Output Specification would include an edited Data Set containing the imputed values, plus a report detailing which values had been imputed.

20. The Process Design specifies the control logic, that is the sequencing and conditional flow logic among different sub-processes (Process Steps). This flow is described in the Process Control Design. When creating a Process Design, a Process Control Design that provides information on "what should happen next" is specified. Sometimes one Process Step will be followed by the same step under all circumstances. In such cases the Process Control Design simply records what Process Step comes next. However, sometimes there will be a choice of which Process Step will be executed next. In this case, the Process Control Design will detail the set of possible "next steps" and the criteria to be applied in order to identify which Process Step(s) should be performed next.

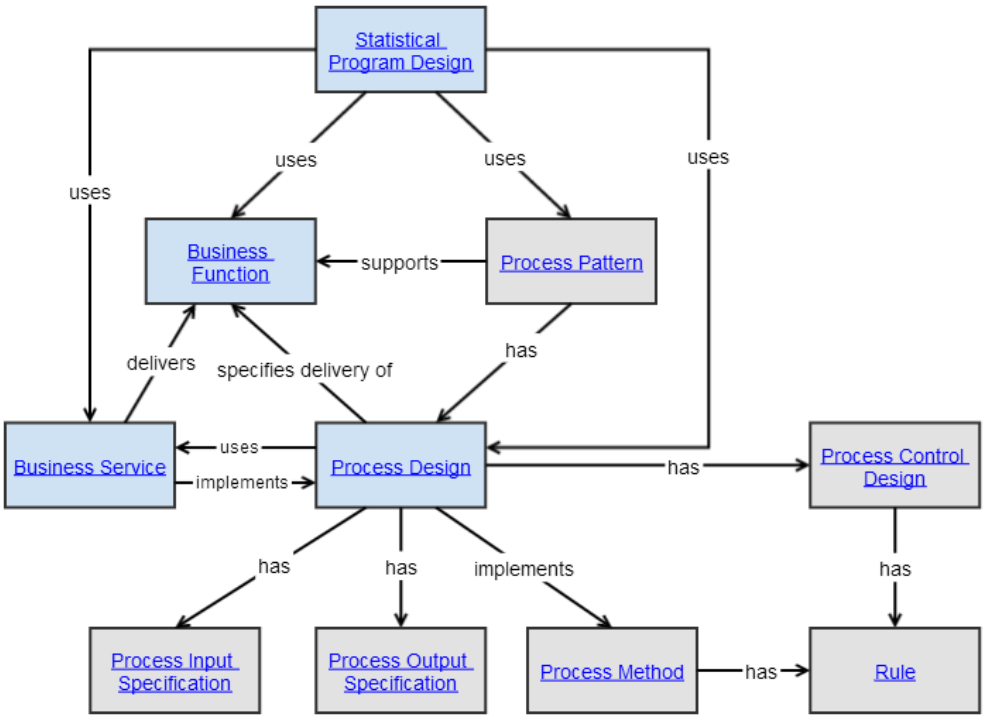
21. The Process Design associated with that Process Step will identify the Process Method that will be used to perform the Business Function associated with the Process Step. For example, if the Business Function is 'impute missing values in the data', the Process Method might be 'nearest neighbour imputation'.

22. A Process Method specifies the method to be used, and is associated with a set of Rules to be applied. For example, any use of the Process Method 'nearest neighbour imputation' will be associated with a (parameterized) Rule for determining the 'nearest neighbour'. In that example the Rule will be mathematical (for example, based on a formula). Rules can also be logical (for example, if Condition 1 is 'false' and Condition 2 is 'false' then set the 'requires imputation' flag to 'true', else set the 'requires imputation flag' to 'false').

23. The resulting Process Design and Process Control Design objects (along with related Process Input Specifications and Process Output Specifications) would be used in the implementation of the Process Step.

B. Using Re-usable Business Services

Figure 5.
Use of re-usable Business Services



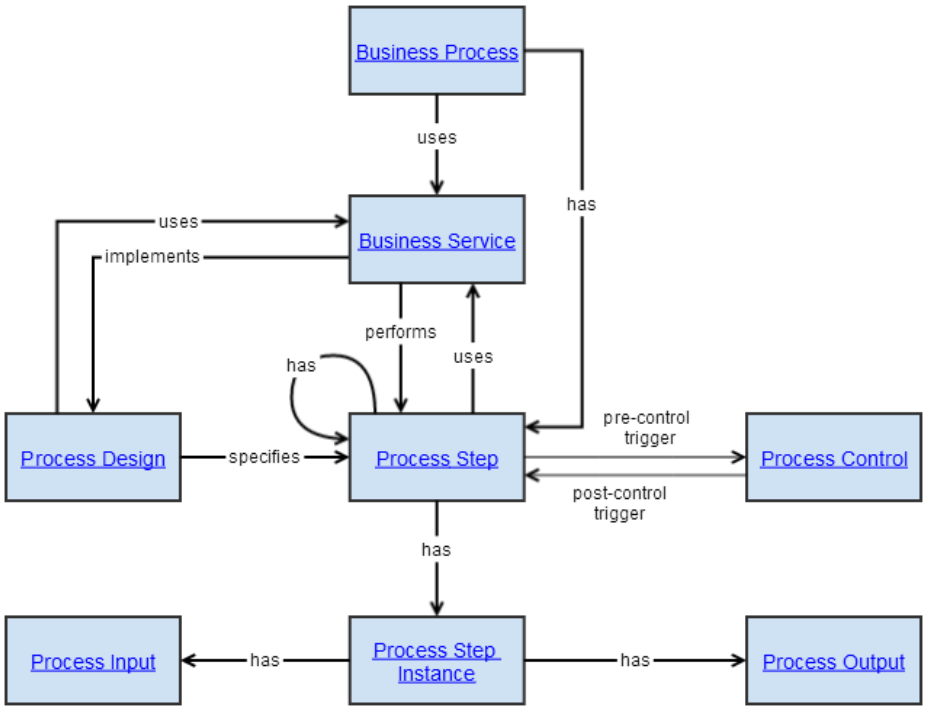
24. It is not always necessary for the Statistical Programme to design its own Process Steps from the beginning. The Common Statistical Production Architecture (CSPA) describes how statistical organizations can create statistical services that are easily reused in other statistical organizations. In GSIM terms, a statistical service is a Business Service. A Business Service is a means of performing a Business Function (an ability that an organization possesses, typically expressed in general and high level terms and requiring a combination of organization, people, processes and technology to achieve).

25. The increased sharing and reuse of Business Services means that the resources needed to meet new demands for statistical production could be considerably reduced, and the time needed to produce new statistical products could be lessened. To facilitate this, CSPA introduced the concept of a statistical services catalogue, where different statistical organizations could list the statistical services they have developed, with the intent of sharing them with other statistical organizations.

26. Business Services have already been designed, with all of the normal input types, output types, process control design, and other properties already specified. Thus, a Business Service can act in a fashion similar to a Process Step designed within the organization, but without the effort required in the traditional scenario.

4. Running Processes

Figure 6.
Run Process



27. A Statistical Programme needs to execute processes to realize some Business Functions. This can be done in two ways: a Process Step can be directly executed by a Business Process, or a re-usable Business Service can be used by the Business Process, as an intermediate trigger for the execution of the Process Step.

28. In order to understand how this works, we characterize the nature of Process Steps in more detail. Process Steps are the resources which have been specified in a Process Design, and which can be executed multiple times. Process Steps can exist at many levels of granularity, and can involve the use of other Process Steps as sub-processes. The navigation

among the sub-processes is performed during execution as indicated by a Process Control, which is itself an implementation of a Process Control Design.

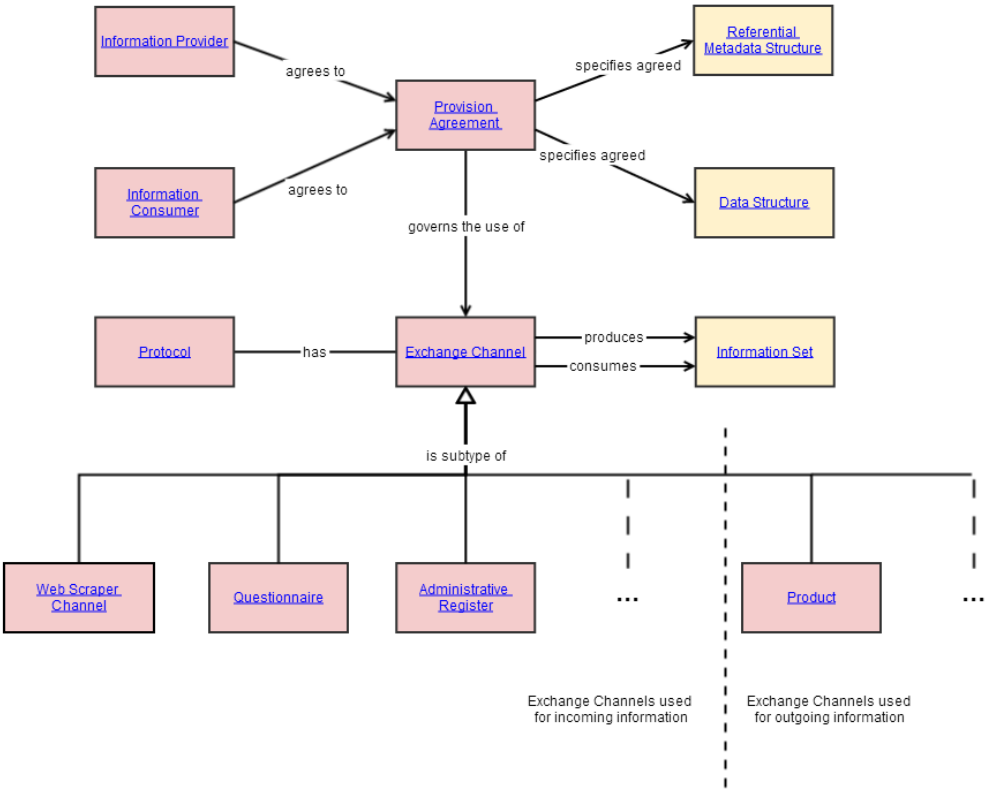
29. Individual executions of a Process Step are represented by the Process Step Instance. It is at this level that specific instances of the inputs and outputs are used. In the Process Design, the types of inputs and outputs are specified (Process Input Specification and Process Output Specification) - the actual instances of inputs and outputs are associated with the Process Step Instance, and are represented by the Process Input and Process Output objects. Inputs can be of any type of information - rules, parameters, data sets, metadata of many kinds, etc. Outputs are similarly of many different types, and often include process metrics and various types of reports, as well as data and metadata.

30. At the time the Process Design is executed someone or something needs to apply the designated method and rules. The Process Design can designate the Business Service that will implement the Process Method at the time of execution. A Business Service represents a service delivered by a piece of software (as described in the section above) or a person. Putting a publication on the statistical institute's website or putting collected response forms in a shared data source for further processing are both examples of Business Services.

31. It should be noted that this model supports both automated and manual processes, and processes which might involve sub-processes of either type.

D. Exchanging Information

Figure 7.
Exchange Channels



32. Statistics organizations collect data and referential metadata from Information Providers, such as survey respondents and providers of Administrative Registers, and disseminate data to Information Consumers, such as government agencies, businesses and

members of the public. Each of these exchanges of data and referential metadata uses an Exchange Channel, which describes the means to receive (data collection) or send (dissemination) information. Information Providers and Information Consumers can be Organizations or Individuals who are either within or external to the statistical organization.

33. Different Exchange Channels are used for collection and dissemination. Examples of collection Exchange Channels include Questionnaire, Web Scraper Channel and Administrative Register. The only example of a dissemination Exchange Channel currently contained in GSIM is Product. Additional Exchange Channels can be added by organizations depending on their needs.

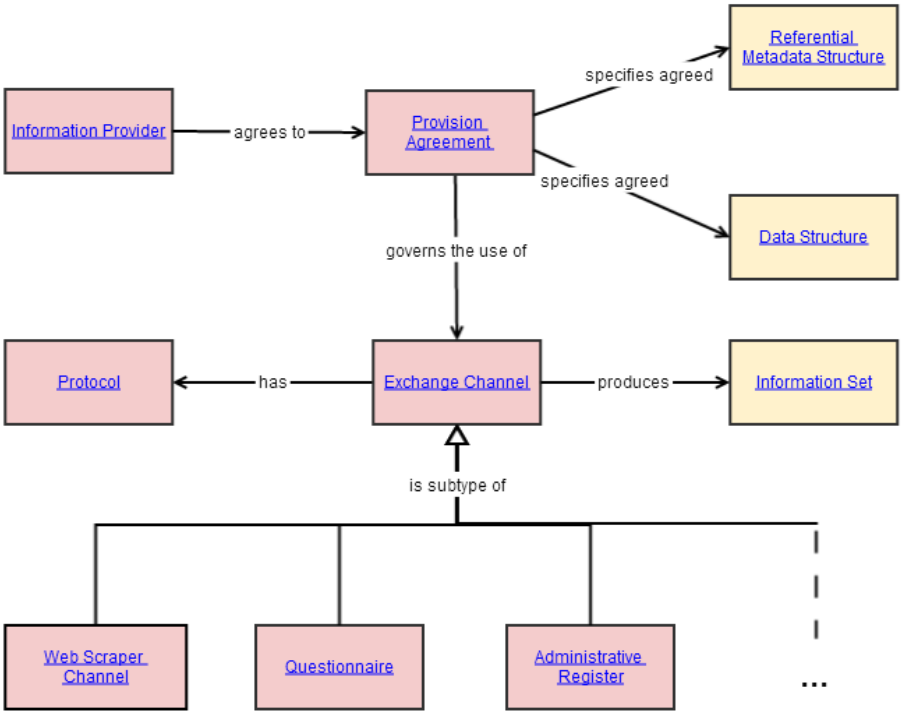
34. The use of an Exchange Channel is governed by a Provision Agreement between the statistics office and the Information Provider (collection) or the Information Consumer (dissemination). The Provision Agreement, which may be explicitly or implicitly agreed, provides the legal or other basis by which the two parties agree to exchange data. The parties also use the Provision Agreement to agree the Data Structure and Referential Metadata Structure of the information to be exchanged.

35. The mechanism for exchanging information through an Exchange Channel is specified by a Protocol (e.g. SDMX web service, data file exchange, face to face interview).

36. To collect data, a statistical organization receives data and referential metadata from the Information Provider in a manner consistent with the Protocol and the Provision Agreement, and the Exchange Channel produces an Information Set. To disseminate data, the Exchange Channel consumes an Information Set, which is then provided to the Information Consumer in a manner consistent with the Protocol and the Provision Agreement. More information about collection and dissemination can be found in the following sections.

5. Collecting Information

Figure 8.
Exchange Channels for collecting information



37. GSIM models three collection Exchange Channel examples: Questionnaire, Web Scraper Channel and Administrative Register. Each of these is detailed in Annex A. Statistics organizations may collect data and referential metadata from Information Providers using additional Exchange Channels, such as file transfer, web services and data scanning. Statistical organizations can extend GSIM to add channels relevant to their context.

38. The use of an Exchange Channel for collection is governed by a Provision Agreement between the statistical organization and the Information Provider. The two parties use the Provision Agreement to agree the Data Structure and Referential Metadata Structure of the data to be exchanged. The mechanism for collecting information through the Exchange Channel is specified by a Protocol (e.g. face to face interview, data file exchange, web robot). The collecting organization uses the collected information to produce an Information Set, which may contain data or referential metadata.

7. Processing and Analysing Information

39. GSIM is very flexible in describing the processing and analysis of information.

40. One can understand the statistical production process from a data-centric perspective[8]. Statistical organizations strive to produce high-quality accurate data that is supported by the metadata needed to make the data optimally useful. For this reason, it is appropriate to think of the evolution of data as it passes through the production process. The focus of many activities is driven by the metadata, but at the end of the production process the metadata is a supporting resource from the perspective of the data and ultimately a statistical product. The relationship of the data and metadata is one which is important to understand.

41. Collected data comes into a statistical organization through an Exchange Channel. Regardless of how the data is collected and where it comes from, it is a resource which will begin a process of evolution through many different stages. The initial data is described as a Data Set with relevant Data Structures. Data Sets are stored in an organised way in a Data Resource. The Data Sets are the primary inputs and outputs of a set of Process Steps, as conducted by a Statistical Program.

42. As the statistical organization moves from raw input data to an increasingly refined set of data, it can be understood that each phase of this processing adds additional Datasets to the Data Resource. There are many different Process Methods which may inform these activities. These are implemented through the different Process Steps that the statistical organization undertakes.

43. At a certain point (and this can take place at different places within the production process, depending on the type of edits being performed) the data will be analysed for the production of statistical Products. The analysis of the data can be understood as using Data Sets from the Data Resource as inputs to processes such as confidentiality routines or to produce explanations of the data. The operations performed during analysis will vary based on what the ultimate Products are - confidentialised unit-record data may be a Product, or we may be publishing aggregated indicators and tables to address specific policy issues, and these involve different types of analysis - but the process is still one of further evolving the information held in the Data Resource.

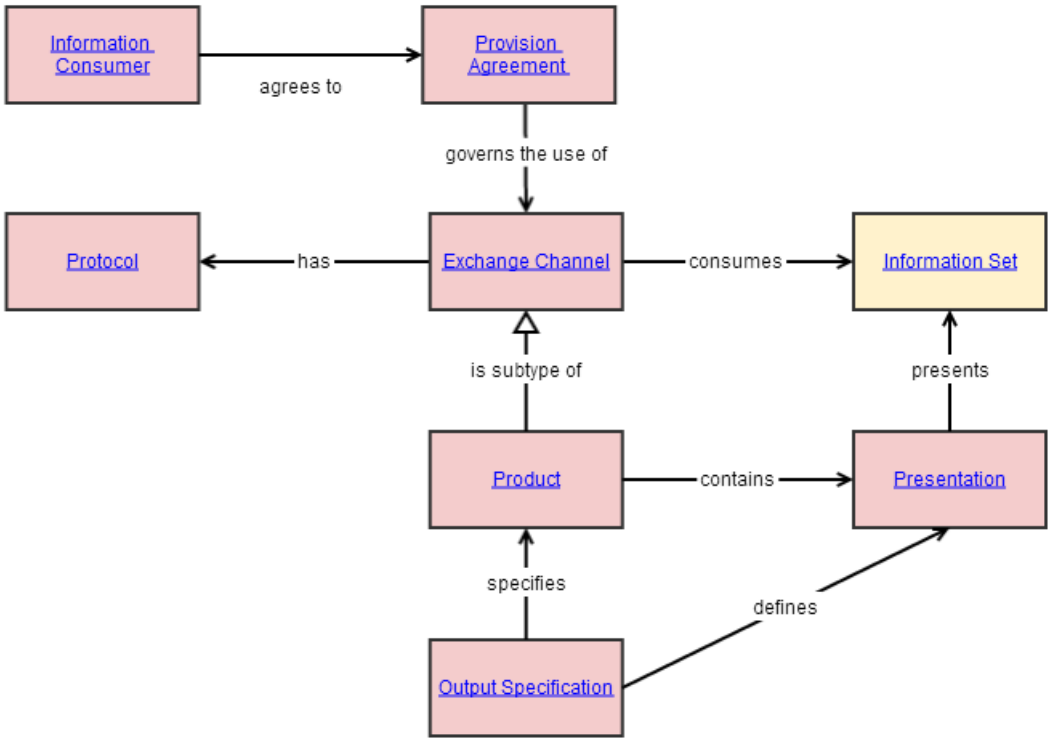
44. In the past, there was an assumption that a data collection will be followed by processing, analysis, and dissemination of the statistical Products. This is a time-consuming and resource-intensive process. One way to make the functions of a statistical organization

[8] Although these paragraphs focus of data, the same descriptions can be applied to referential metadata.

more efficient is to re-use data to produce new Products as they are asked for, lowering the cost and shortening the time needed for production. In this sense, the Data Resource can be understood as an organizational asset, to be managed and exploited to the greatest extent possible.

8. Disseminating Information

Figure 9.
Exchange Channel for disseminating information



45. A statistical organization disseminates statistical information to an Information Consumer.

46. The Information Consumer accesses a set of information via a Product (or potentially via another Exchange Channel), which contains one or more Presentations. Each Presentation will typically provide a view of data and associated metadata to define and describe the structure of the presented data, and perhaps referential metadata in the form of textual media, such as quality reports.

47. A Presentation can take different forms - for example, it could be a screen visualization of a table of data in graphical form displayed in an HTML page, a downloadable PDF, or an SDMX file in XML format.

48. An Output Specification defines what is contained in the Presentation. A Product, which packages Presentations, may be a statistical organization's standard specific output as one might see in:

- a regular statistical bulletin (e.g. a monthly publication of the Retail Prices Index),
- a dynamically generated package of statistical content which is generated following the receipt of a query from an Information Consumer who wishes to access the organization's data via a published API (Application Programming Interface) or

- some data exploration facility which might be built into the statistical organization's website.

49. The Output Specification also defines the information required from the Information Set for the Presentation. The specifications are frequently determined by an internal (to the organization) process which would have specifically standard, static outputs to produce (such as the aforementioned statistical bulletins). For dynamically delivered products, aspects of the specification could be determined by the Information Consumer at run time, via machine to machine dynamic, as exemplified in the API scenario above. In either case, the requests would result in the Output Specification specifying Information Set data and/or referential metadata that will be included in each Presentation.

50. The mechanism for providing a Product is specified by a Protocol (e.g. SDMX-ML, DDI XML, PDF etc.). This formatting information forms part of the Output Specification to generate the Product and its Presentations in the appropriate format.

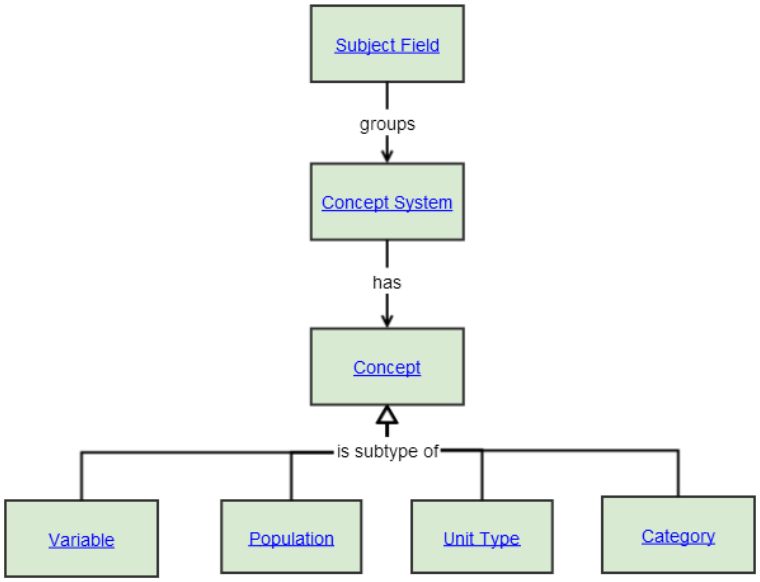
51. The Information Consumer can be one of many forms depending upon the scenario of the request. The Information Consumer could be a person accessing the statistical organization's website and visually inspecting the contents of a web page, or it could be a computer program requesting the information via an API using an SDMX query. The Information Consumer's access to the information would be subject to a Provision Agreement, which would set out the conditions of access and use. This might be in the form of passive acceptance of the terms and conditions of use of the data from a website the Information Consumer is accessing, or in the case of access to a greater level of detail via an API, it might be a more involved registration process.

B. Foundational Information

52. The GSIM Concepts and Structures groups include information objects which are foundational to the statistical Business Process. That is, these objects are the conceptual and structural objects which are used as the Process Inputs and Process Outputs to the process. The Concepts area of GSIM includes the sets of information objects that describe and define the terms used when talking about the real-world phenomena that the statistics measure in their practical implementation. The Structures area includes the set of information objects used in relation to data and referential metadata and their structures. The objects described in this section of the document are used to provide information that helps users of data and metadata understand the results of Business Processes and Statistical Programmes.

1. Concepts

Figure 10.
Concepts



53. At an abstract level, a Concept is defined in GSIM as a 'unit of thought differentiated by characteristics'. Concepts are used in different ways throughout the statistical lifecycle, and each different role of a Concept is described using a different information object (the objects are subtypes of Concept). A Concept can be used in these situations:

(a) As a characteristic. The Concept is used by a Variable to describe the particular characteristic that is to be measured about a Population. For example, to measure the Concept of gender in a population of adults in the Netherlands, the Variable combines this Concept with the Unit Type person.

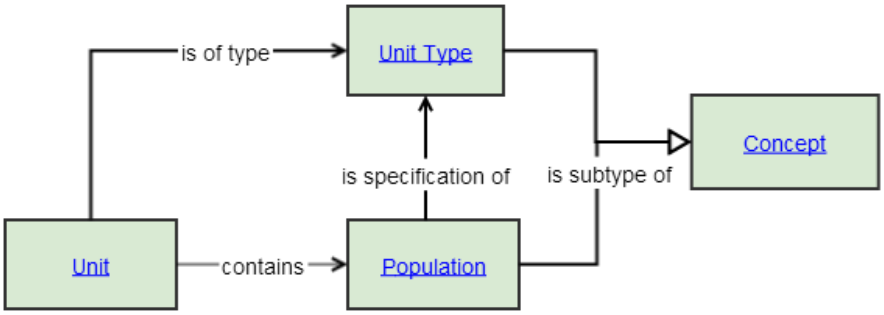
(b) As a Unit Type or a Population. To describe the set of objects that information is to be obtained about in a statistical survey. For example, the Population of adults in the Netherlands, based on the Unit Type of persons.

(c) As a Category to further define details about a Concept. For example, Male and Female for the Concept of Gender. Codes can be linked to a Category via a Node (i.e., a Code Item or Classification Item), for use within a Code List or Statistical Classification.

54. Concept Systems are sets of Concepts which are structured by the relations between those Concepts. A Subject Field groups Concept Systems on the basis of their field of special knowledge (for example, labour market, tourism).

2. Population

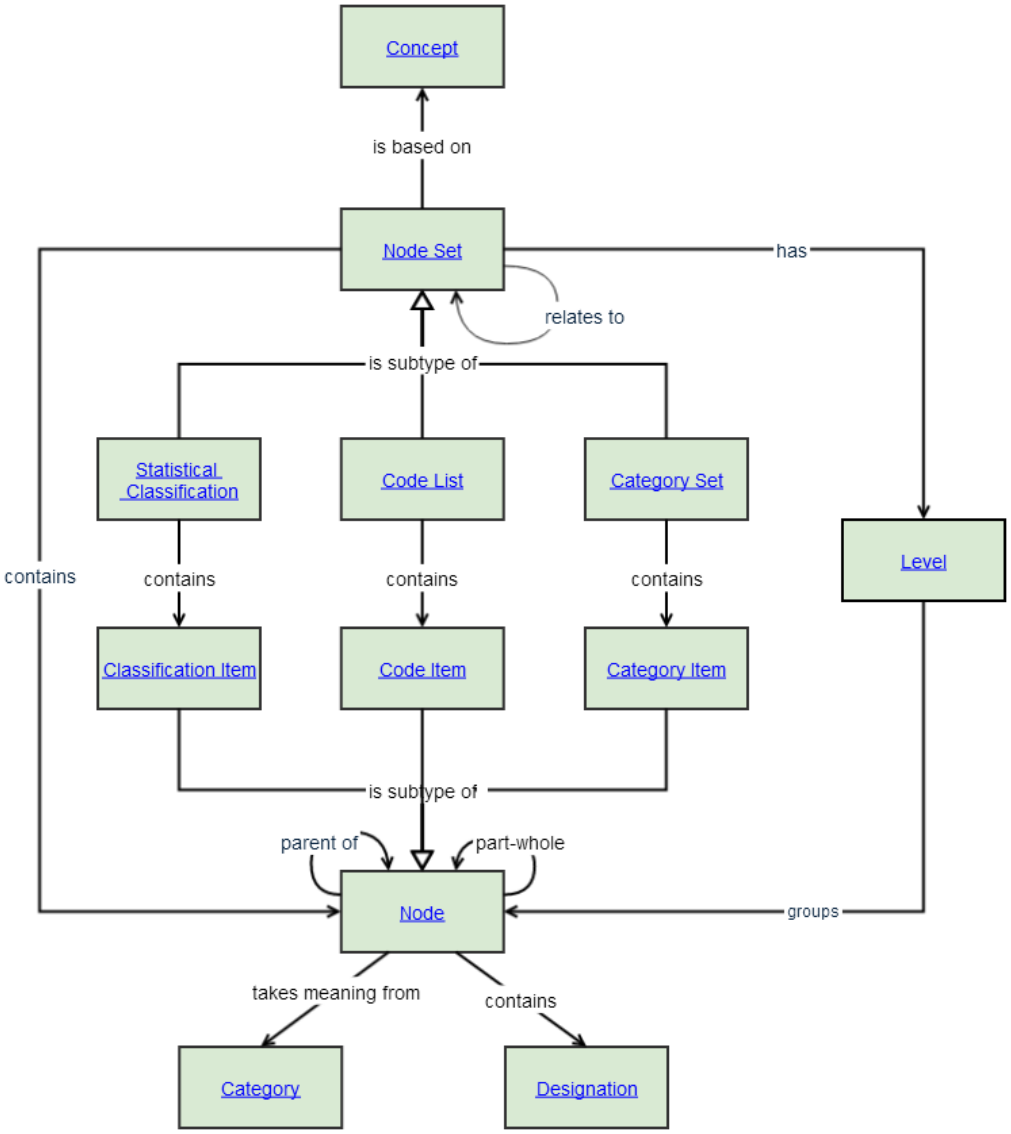
Figure 11.
Populations and Units



55. There are several kinds of Populations depending on what Process Step it is used in. For example a statistical organization may refer to a target, survey, frame, or analysis population. The objects of interest in a statistical process are Units (for example, a particular person or a business). Data are collected about Units. There are two ways in which a unit is specified in the model. A Unit is an individual entity associated with a Population about which information may be obtained. A Unit Type (for example persons or businesses) is a way of identifying an abstract type of Unit that a Variable is measuring.

3. Node and Node Set

Figure 12.
Node and Node Set inheritance



56. A Category is a particular type of Concept whose role is to define a characteristic. There are three ways in which a Category can be used. In GSIM, these are described as the three subtypes of Node - Category Item, Code Item and Classification Item. Categories are grouped into Node Sets based on the way in which it can be used. There are three subtypes of these groups (Node Sets) - Category Sets, Code Lists and Statistical Classifications.

57. A Category Set is a set of Category Items, which contain the meaning of a Category without any associated representations. An example of a Category Set is: Male, Female.

58. In a Code List, the Code Items contain the meaning of the Categories combined with a Code representation. An example of a Code List is: 1. Male, 2. Female.

59. A Statistical Classification is similar to a Code List. It combines the meaning of the Category with a Code representation. However the content of a Statistical Classification must fulfil certain criteria and have a certain status. The Classification Items must be

mutually exclusive and jointly exhaustive for the Level at which they exist at in the Statistical Classification. An example of a Statistical Classification is: 1. Male, 2. Female, 3. Intersex.

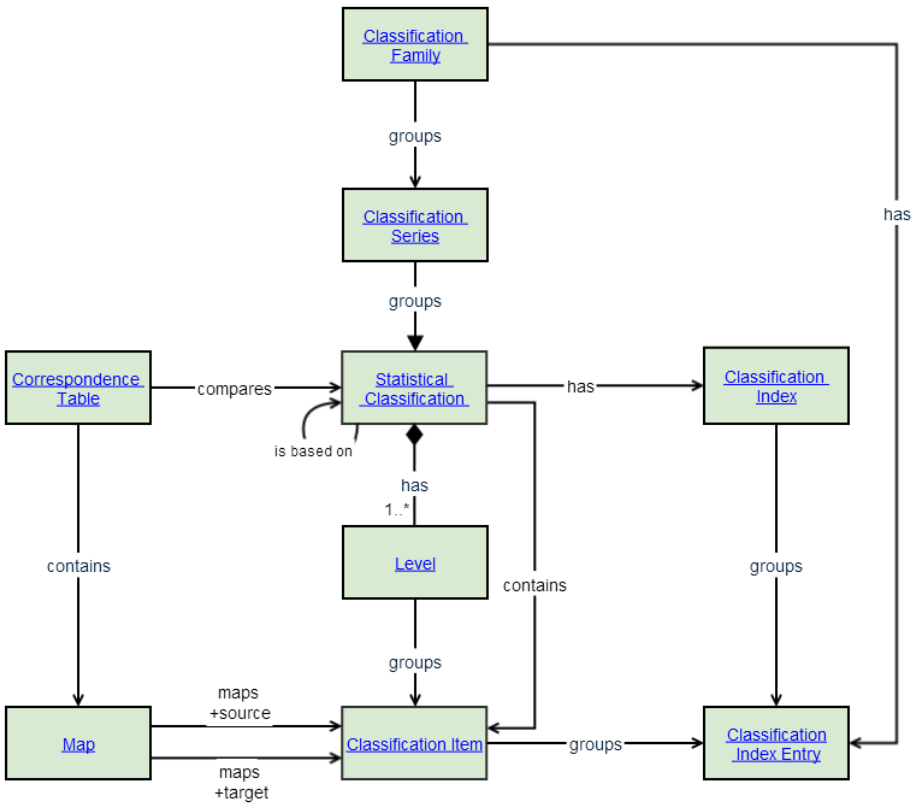
60. A Code List does not have to satisfy the same criteria as the Statistical Classification. The Code List can also contain additional Code Items to support a particular use of the Code List, such as the inclusion of missing values.

61. The similarities between Statistical Classifications, Code Lists and Category Sets are inherent through their link (as subtypes) to Node Set. Similarly, the three types of item which make up each group (Classification Item, Code Item and Category Item respectively) are subtypes of Node.

4. Statistical Classification

62. This section describes a Statistical Classification and its related management objects, as a particular view of the Node Set portion of GSIM. Further detail about Statistical Classifications in particular can be found in the GSIM Statistical Classification Model.

Figure 13.
Statistical Classifications



63. The figure above provides an overview of the objects relating to Statistical Classifications.

64. A Classification Family is a group of Classification Series related based on a common Concept (e.g. economic activity). A Classification Series is an ensemble of one or more Statistical Classifications that are based on the same Concept. The Statistical Classifications in a Classification Series are related to each other as versions or updates. Typically, these Statistical Classifications have the same name, for example International

Standard Industrial Classification of All Economic Activities (ISIC), or International Standard Industrial Classification of Occupations (ISCO).

65. A Statistical Classification is a set of Categories which may be assigned to one or more Represented Variables used in the production and dissemination of statistics. The Categories at each Level of the classification structure must be mutually exclusive and jointly exhaustive of all objects/units in the population of interest. One example of a Statistical Classification is ISIC rev 4.

66. The Categories are defined to reference one or more characteristics of a particular population of interest. A Statistical Classification may have a flat, linear structure or may be hierarchically structured, such that all Categories at lower Levels are sub-categories of a Category at the next Level up.

67. A Statistical Classification has Categories that are represented by Classification Items. These Classification Items are organised into Levels determined by the hierarchy. A Level is a set of Concepts that are mutually exclusive and jointly exhaustive; for example: section, division, group and class in ISIC rev 4.

68. A Classification Item combines the meaning from a Category, its representation (i.e., Code) and additional information in order to meet the Statistical Classification criteria, for example "A- agriculture, forestry and fishing" and accompanying explanatory text such as information about what is included and excluded.

69. Statistical Classifications can be versions or variants. A variant type of Statistical Classification is based on a version type of Statistical Classification. In a variant the Categories of the version may be split, aggregated or regrouped to provide additions or alternatives to the standard order and structure of the original Statistical Classification.

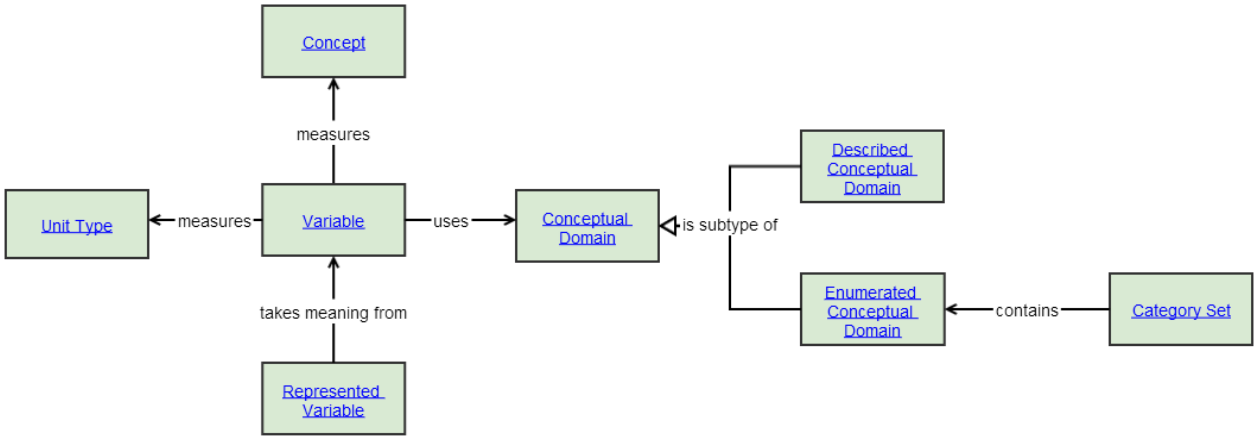
70. A Correspondence Table is a set of Maps. These Maps link a Classification Item in a Statistical Classification with a corresponding Classification Item in another Statistical Classification via the Concept which is common to both Classification Items. For example, in a Correspondence Table displaying the relationship between ISIC rev 4 and the North American Industry Classification System (NAICS 2007 (US)), "0112 - Growing of Rice" in ISIC Rev 4 is related to "111160 - Rice Farming" in NAICS through the common concept of "growing rice".

71. A Classification Index shows the relationship between text found in statistical data sources (responses to survey questionnaires, administrative records) and one or more Statistical Classifications. A Classification Index may be used to assign the Codes for Classification Items to observations in Statistical Programs.

72. A Classification Index Entry is a word or short text (e.g. the name of a locality, an economic activity or an occupational title) describing a type of Concept to which a Classification Item applies, together with the Code of the corresponding Classification Item. Each Classification Index Entry typically refers to one item of the Statistical Classification. Although a Classification Index Entry may be associated with a Classification Item at any Level of a Statistical Classification, they are normally associated with Classification Items at the lowest Level.

5. Variable

Figure 14.
Variable



73. When used as part of a Business Process, a Unit Type defining a Population is associated with a characteristic. The association of Unit Type and a Concept playing the role of a characteristic is called a Variable (see Figure 14). For example, if the Population is adults in Netherlands, then a relevant Variable might be the Concept educational attainment combined with the Unit Type person.

74. The Variable (person's educational attainment) does not include any information on how the resulting value may be represented. This information (the Value Domain) is associated with the Represented Variable. This distinction promotes the reuse of a Variable definition when what is being measured is conceptually the same but it is represented in a different manner.

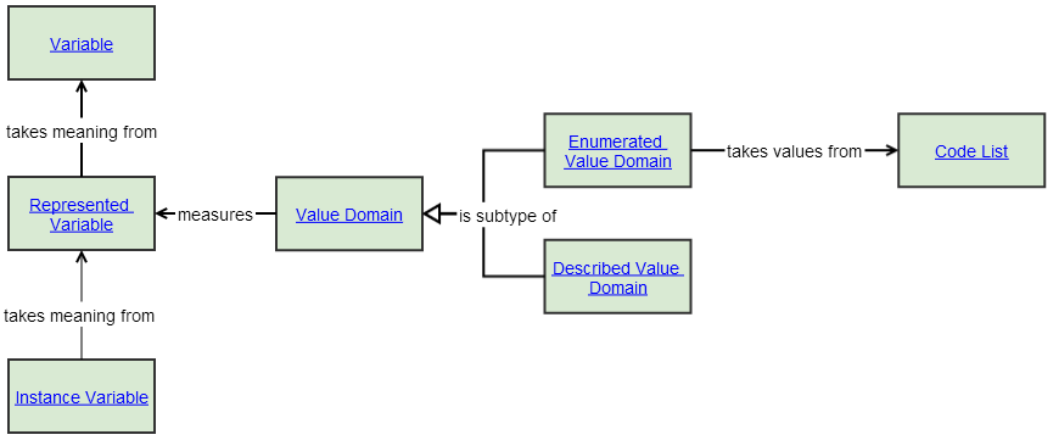
75. A derived variable is created by a Process Step that applies a Process Method to one or more Process Inputs (Variables). The Process Output of the Process Step is the derived variable.

76. A Conceptual Domain is associated with a Variable. It has two subtypes: Described Conceptual Domain and Enumerated Conceptual Domain. An Enumerated Conceptual Domain, in combination with a Category Set, contains information on the semantics of the Categories used by the Variable.

6. Represented Variable

77. GSIM assists users in understanding both the meaning and the concrete data-representation of the object. Accordingly, GSIM distinguishes between conceptual and representation levels in the model, to differentiate between the objects used to conceptually describe information, and those that are representational.

Figure 15.
Represented Variable



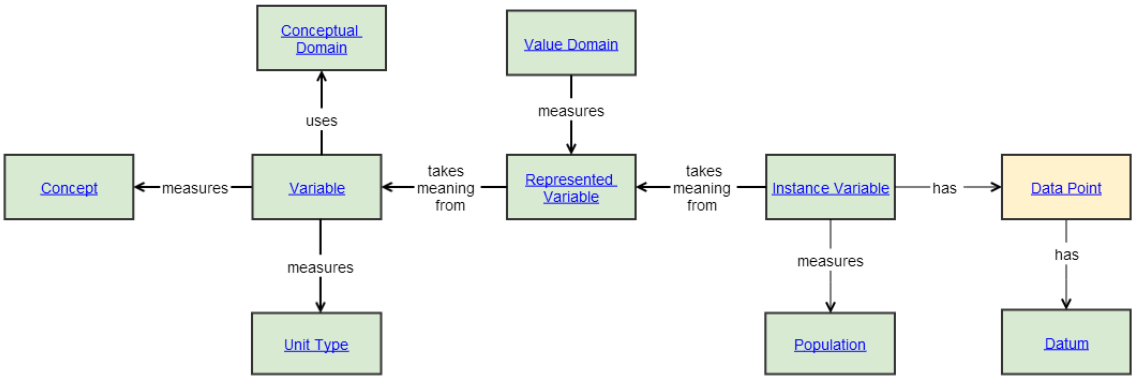
78. The Represented Variable (see Figure 15) adds information that describes how the resulting values may be represented through association with a Value Domain. While Conceptual Domains are associated with a Variable, Value Domains are associated with a Represented Variable. These two domains are distinguished because GSIM separates the semantic aspect (Conceptual Domain) and the representational aspect (Value Domain).

79. Both the Enumerated Value Domain and the Described Value Domain (the two subtypes of Value Domain) give information on how the Represented Variable is represented. The Enumerated Value Domain does this in combination with a Code List, while the Described Value Domain provides a definition of how to form the values, rather than explicitly listing them.

80. The Value Domain includes data type and unit of measure information. The data type contains information on the allowed computations one may perform on the Datum (nominal-, ordinal-, interval-data, etc.), while the unit of measure (Tonnes, Count of __, Dollars, etc.) refines the measure of the Value Domain. For example gender codes lead to nominal statistical data, whereas age values in years lead to interval data.

6. Instance Variable

Figure 16.
Instance Variable



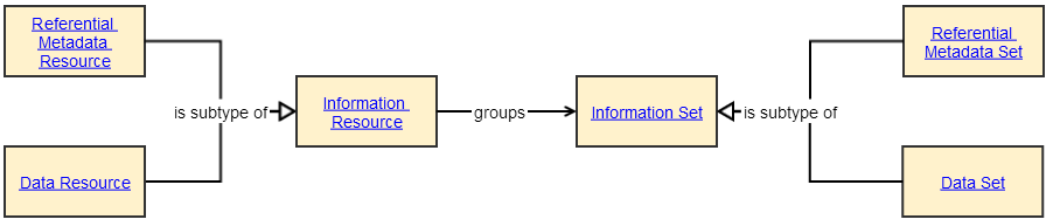
81. An Instance Variable (see Figure 16) is a Represented Variable that has been associated with a Data Set. This can correspond to a column of data in a database. For example, the “age of all the US presidents either now (if they are alive) or the age at their deaths” is a column of data described by an Instance Variable, which is a combination of

the Represented Variable describing "Person's Age" and the Value Domain of "decimal natural numbers (in years)".

82. A Datum is contained within a Data Point in a Data Set. It may be defined by the measure of a Value Domain associated with a describing Instance Variable, combined with the link to a Unit (for unit data), or a Population (for dimensional data).

8. Information Resources

Figure 17.
Information Resources



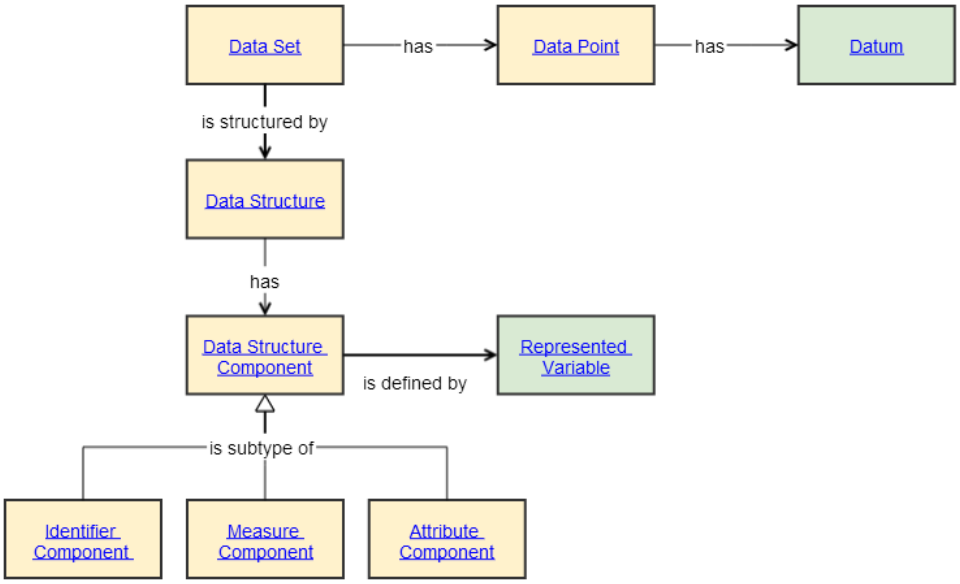
83. Statistical organizations collect, process, analyse and disseminate Information Sets, which are either data (Data Sets) or referential metadata (Referential Metadata Sets).

84. Each Data Set must be structured according to a Data Structure (for example, a structure for Balance of Payments, Demography, Tourism, Education etc.). In the same way, a Referential Metadata Set must be structured according to a Referential Metadata Structure (e.g. an organization's quality framework).

85. Information Resources contain Information Sets. The main purpose of the Information Resource is to aid discovery and management of Information Sets, by providing location and other information relevant to these tasks. There are two types of Information Resource. Data Resources contain Data Sets, and Referential Metadata Resources contain Referential Metadata Sets.

9. Data Sets

Figure 18.
Data Sets



86. A Data Set has Data Points. A Data Point is placeholder (for example, an empty cell in a table) in a Data Set for a Datum. The Datum is the value that populates that placeholder (for example, an item of factual information obtained by measurement or created by a production process). A Data Structure describes the structure of a Data Set by means of Data Structure Components (Identifier Components, Measure Components and Attribute Components). These are all Represented Variables with specific roles.

87. Data Sets come in different forms, for example as Administrative Registers, Time Series, Panel Data, or Survival Data, just to name a few. The type of a Data Set determines the set of specific attributes to be defined, the type of Data Structure required (Unit Data Structure or Dimensional Data Structure), and the methods applicable to the data.

88. For instance, an administrative register is characterized by a Unit Data Structure, with attributes such as its original purpose or the last update date of each record. It contains a record identifying variable, and can be used to define a Population that is used as a frame, to replace or complement existing surveys, or as an auxiliary input to imputation. Record matching is an example of a method specifically relevant for registers.

89. An example for a type of Data Set defined by a Dimensional Data Structure is a time series. It has specific attributes such as frequency and type of temporal aggregation and specific methods, for example, seasonal adjustment, and must contain a temporal variable.

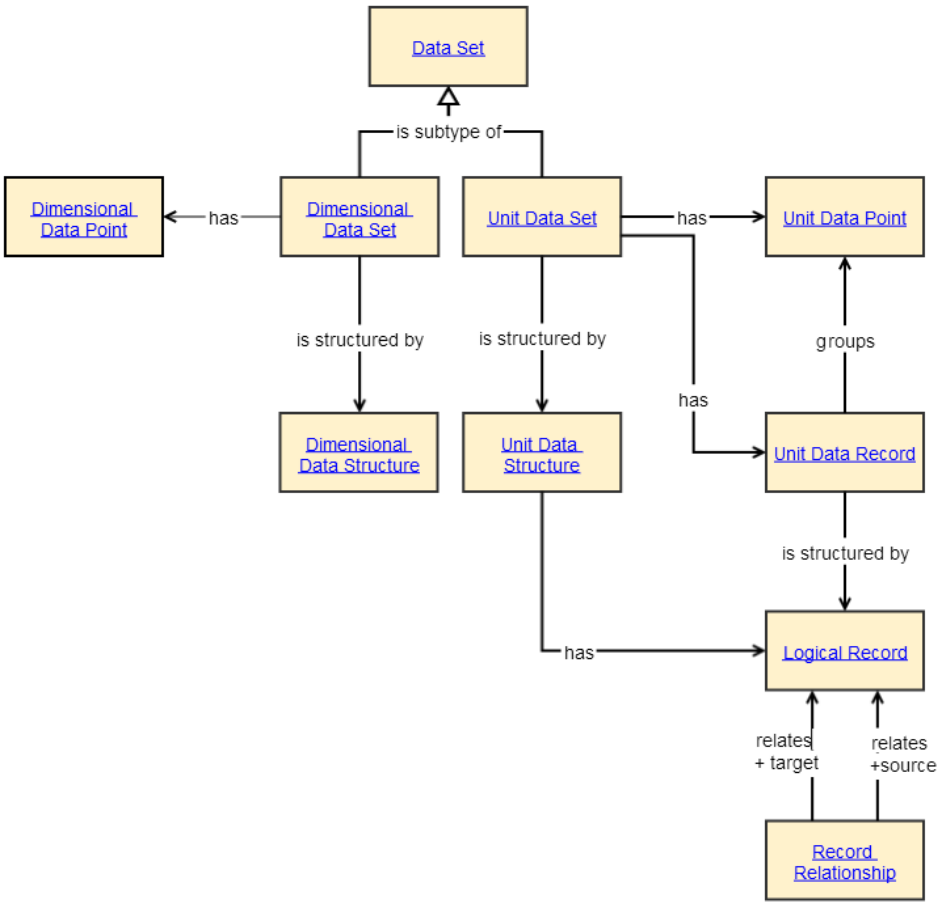
90. Unit data and dimensional data are perspectives on data. Although not typically the case, the same set of data could be described both ways. Sometimes what is considered dimensional data by one organization (for example, a national statistical office) might be considered unit data by another (for example, Eurostat where the unit is the member state). A particular collection of data need not be considered to be intrinsically one or the other. This matter of perspective is conceptual. In GSIM, the distinction is that a Unit Data Set contains data about Units and a Dimensional Data Set contains data about either Units or Populations.

91. GSIM states that all Data Sets must have a structure associated with them. There are, however, cases where a Data Set has no structure – because it was not stored or lost, or it is not known. This type of data may become more prevalent for statistical organizations in the

future. In order for a statistical organization to use this data, the data will need to go through a process of being structured. For example, in a case of investigation of new potential data sources for a new or changed Statistical Need, there will need to be a process where these new data are analysed to determine their content and structure. It is only after this process that these new Data Sets can be described using the Data Structure objects. This unstructured data is currently described by GSIM as a Process Input. Organizations could extend GSIM to capture this use case by creating a new subtype of the Information Set object.

10. Dimensional and Unit Data Structures

Figure 19.
Data Structures



92. A Dimensional Data Structure describes the structure of a Dimensional Data Set by means of Represented Variables with specific roles.

93. The combination of dimensions contained in a Dimensional Data Structure creates a key or identifier of the measured values. For instance, country, indicator, measurement unit, frequency, and time dimensions together identify the cells in a cross-country time series with multiple indicators (for example, gross domestic product, gross domestic debt) measured in different units (for example, various currencies, per cent changes) and at different frequencies (for example, annual, quarterly). The cells in such a multi-dimensional table contain the observation values.

94. A measure is the variable that provides a container for these observation values. It takes its semantics from a subset of the dimensions of the Dimensional Data Structure. In

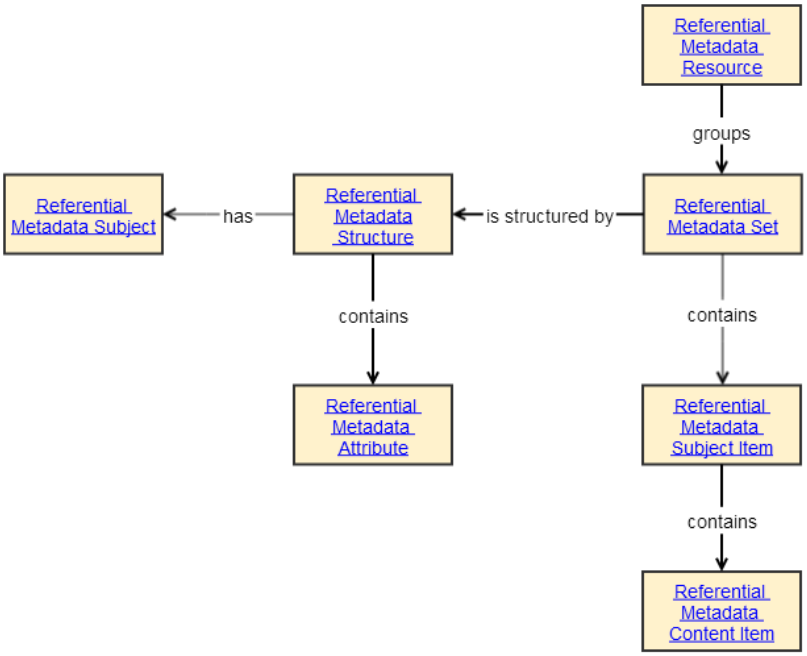
the previous example, indicator and measurement unit can be considered as those semantics-providing dimensions, whereas frequency and time are the temporal dimensions and country the geographic dimension. An example for a measure in addition to the plain 'observation value' could be 'pre-break observation value' in the case of a time series. Dimensions typically refer to Represented Variables with coded Value Domains (Enumerated Value Domains), measures to Represented Variables with uncoded Value Domains (Described Value Domains).

95. A Unit Data Structure describes the structure of a Unit Data Set by means of Represented Variables with specific roles. It distinguishes between the logical and physical structure of a Data Set. A Unit Data Set may contain data on more than one type of Unit, each represented by its own record type.

96. Logical Records describe the structure of such record types, independent of physical features by referring to Represented Variables that may include a unit identification (for example, household number). A Record Relationship defines source-target relations between Logical Records.

11. Referential Metadata Sets

Figure 20.
Referential Metadata Sets



97. Information that describes the characteristics of statistics is “referential metadata”. These metadata can be broad, such as about an entire Statistical Program, or narrow, such as about an individual Data Point. Referential Metadata Resources, a special type of Information Resource, provide top-level containers for referential metadata.

98. A Referential Metadata Set organizes referential metadata, whose structure is defined in a Referential Metadata Structure. A Referential Metadata Structure specifies both the Referential Metadata Subject for which referential metadata may be included, and a structured list of Referential Metadata Attributes that can be reported or authored for the given Referential Metadata Subject.

99. These subjects may be any GSIM object type, or any Data Point or set of Data Points created from a specific Data Structure.

- Example of a GSIM object type as a Referential Metadata Subject: Product for which there is a list specified in a Value Domain. The Value Domain specifies the list of actual Products for which reference metadata can be reported or authored using this Referential Metadata Structure.
- Examples of Referential Metadata Attributes include status, coverage, methodology description, and quality indicator.

100. A Referential Metadata Set contains the actual referential metadata reported or authored. The Referential Metadata Subject Item identifies the actual object e.g. actual Product such as Balance of Payments and International Investment Position, Australia, June 2013, or actual Data Points such as the Data Points for a single region within a Data Set covering all regions for a country.

101. The Referential Metadata Content Item is the actual metadata for the identified Referential Metadata Subject Item. Each Referential Metadata Content Item contains the reported referential metadata for one Referential Metadata Attribute specified in the Referential Metadata Structure.

Table 1.
Example of Use of GSIM Referential Metadata Objects

GSIM Object	ONS Statistical bulletin: Public Sector Finances, October 2013:**Table 1**
Referential Metadata Structure	Implicit
Referential Metadata Subject	Data Structure Component
Referential Metadata Attribute	Table footnote
Referential Metadata Set	Footnotes
Referential Metadata Subject Item	*Data Structure Component*: £billion; PS Current Budget; PS Current Budget ex APF;
Referential Metadata Content Item	Footnoted text