II. Information in the Statistical Business Process

8.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.

A. Identifying and Evaluating Statistical Needs

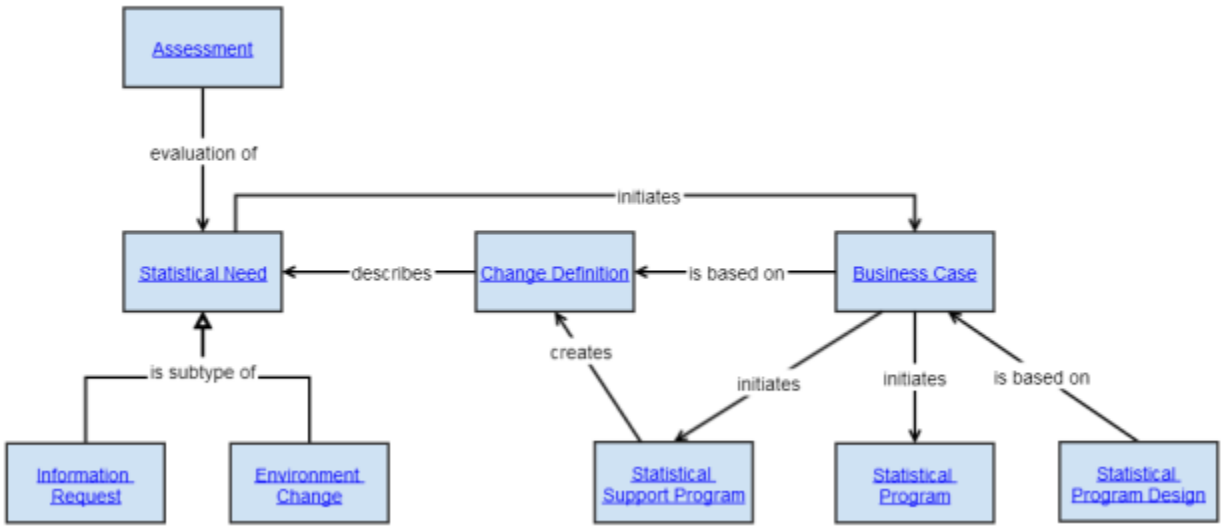

Figure 1. Identify and Evaluate Statistical Needs

9. 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*.

10. *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.

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

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

13. 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.

B: Designing and Managing Statistical Programs

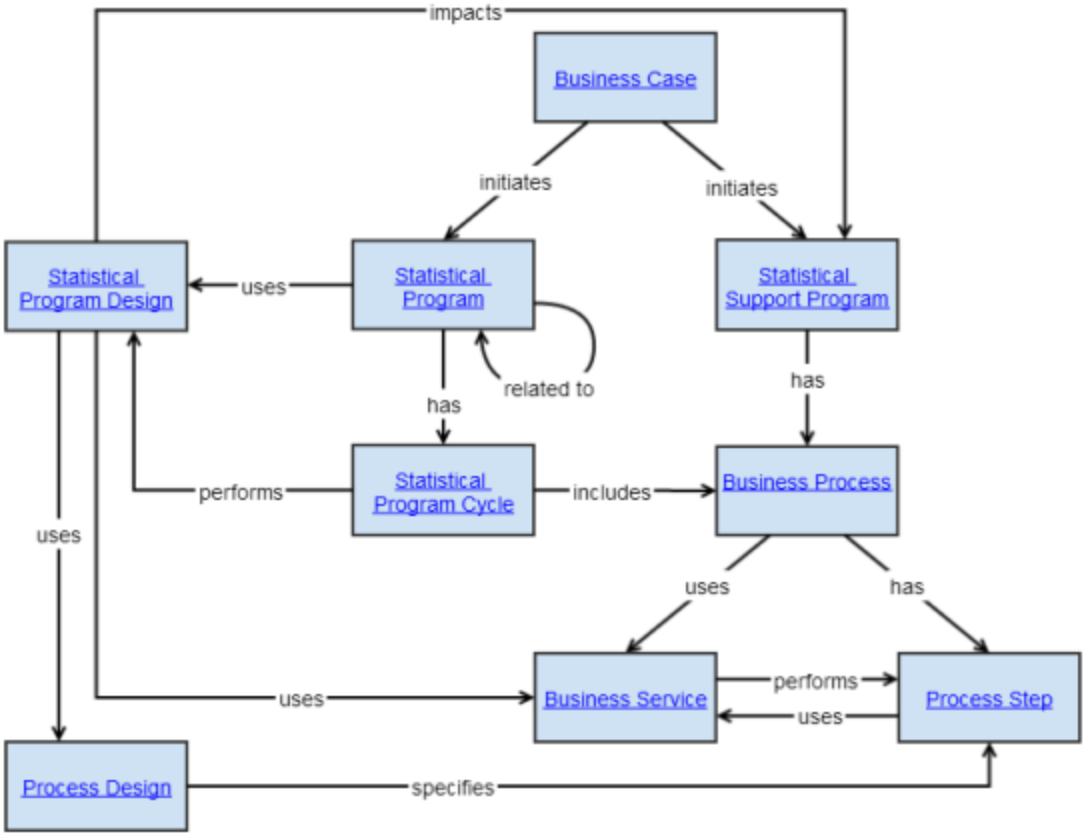

Figure 2. Design and Manage Statistical Program

14. 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 Program*, the creation of a new *Statistical Program*, or the evolution of an existing *Statistical Program Design* to be implemented by an existing *Statistical Program*.

15. *Statistical Support Programs* undertake the activities of the statistical organization such as statistical change programs, data management programs, metadata management programs, methodological research programs, etc. A good example is a program which manages classifications.

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

17. *Statistical Programs* require *Statistical Program 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 Program 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 Program* might perform three iterations of data collection and processing, then analyze the data and disseminate the resulting statistical *Products*. Each of these activities could be understood to be a separate *Business Process*.

18. The *Statistical Program 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 Program Design*.)

19. It should be noted that *Statistical Program 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.

C: Designing Process Steps

20. 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 P*roc ess Step* can be as "large scale" or "small scale" as the designer of a particular *Business Process* chooses (see Figure below).

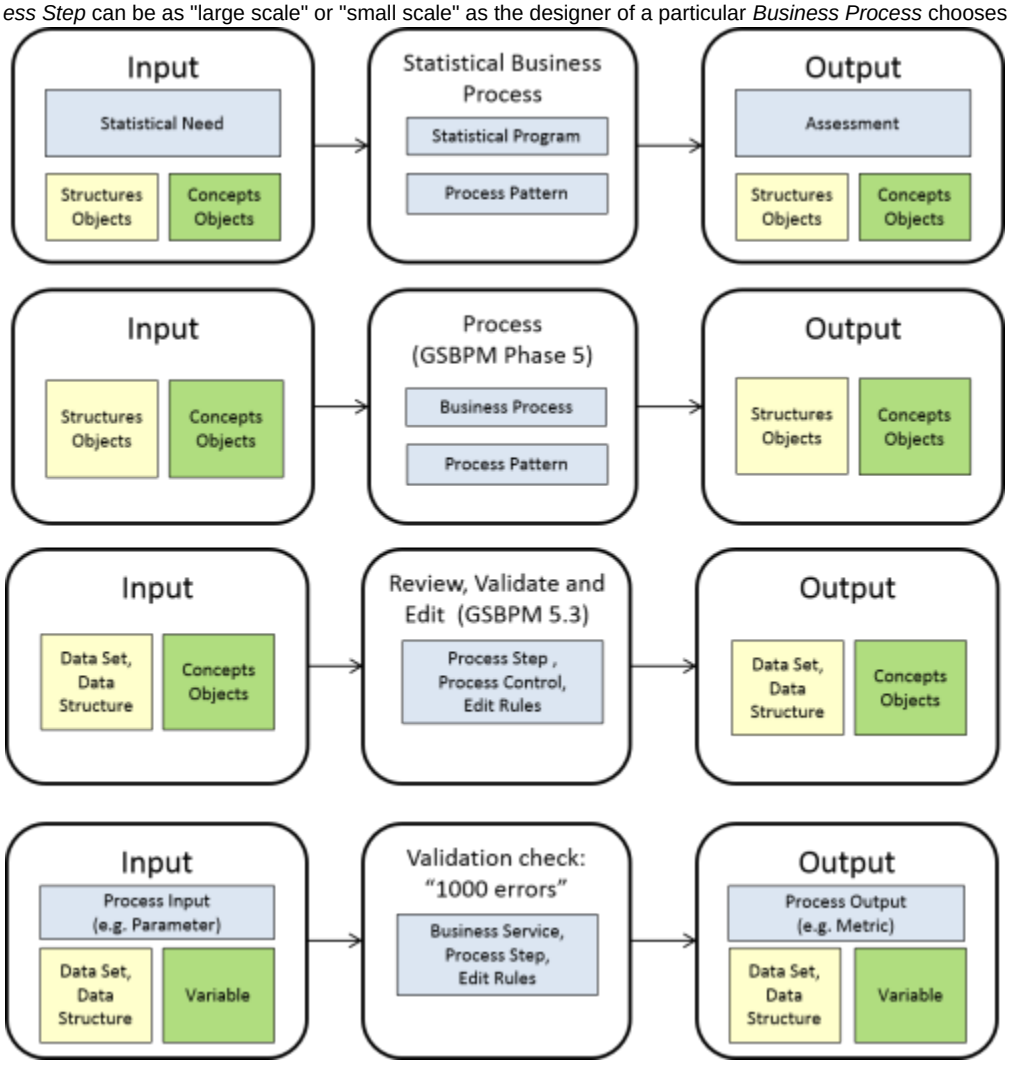

Figure 3. Process Steps can be as large or small as needed

21. *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*.

22. 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.

i. Designing *Process Steps*

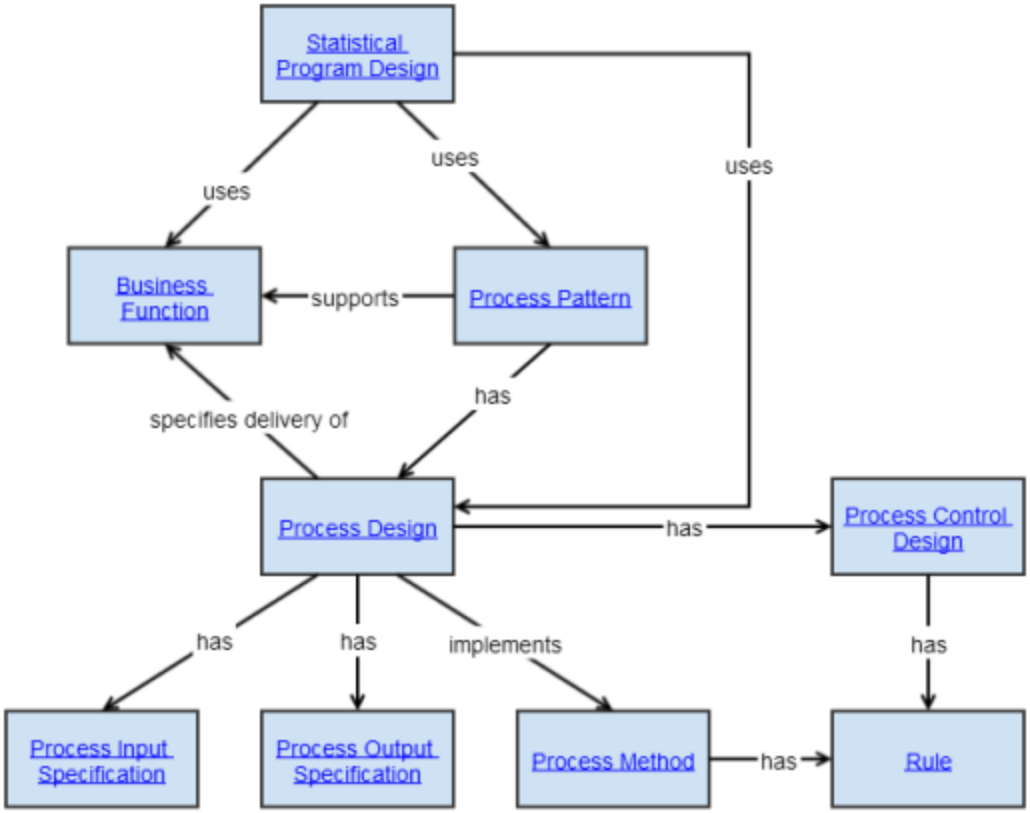

Figure 4. Design Process Steps

23. A *Statistical Program 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 Program* 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*.

24. 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.

25. *Process Designs* specify several things: they identify the different types of inputs and outputs represented by the *Process Input Specification* a nd *Process Output Specification*. Examples of *Process Inputs* include data, metadata such as *Statistical Classifications*, imputation and editing *Ru les*, 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.

26. 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.

27. 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 *Proce ss 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.

28. The *Process Design* associated with that *Process Step* will identify the *Process Method* that will be used to perform the *Business Function* ass ociated 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'.

29. 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').

30. 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*.

ii. Using Re-Usable *Business Services*

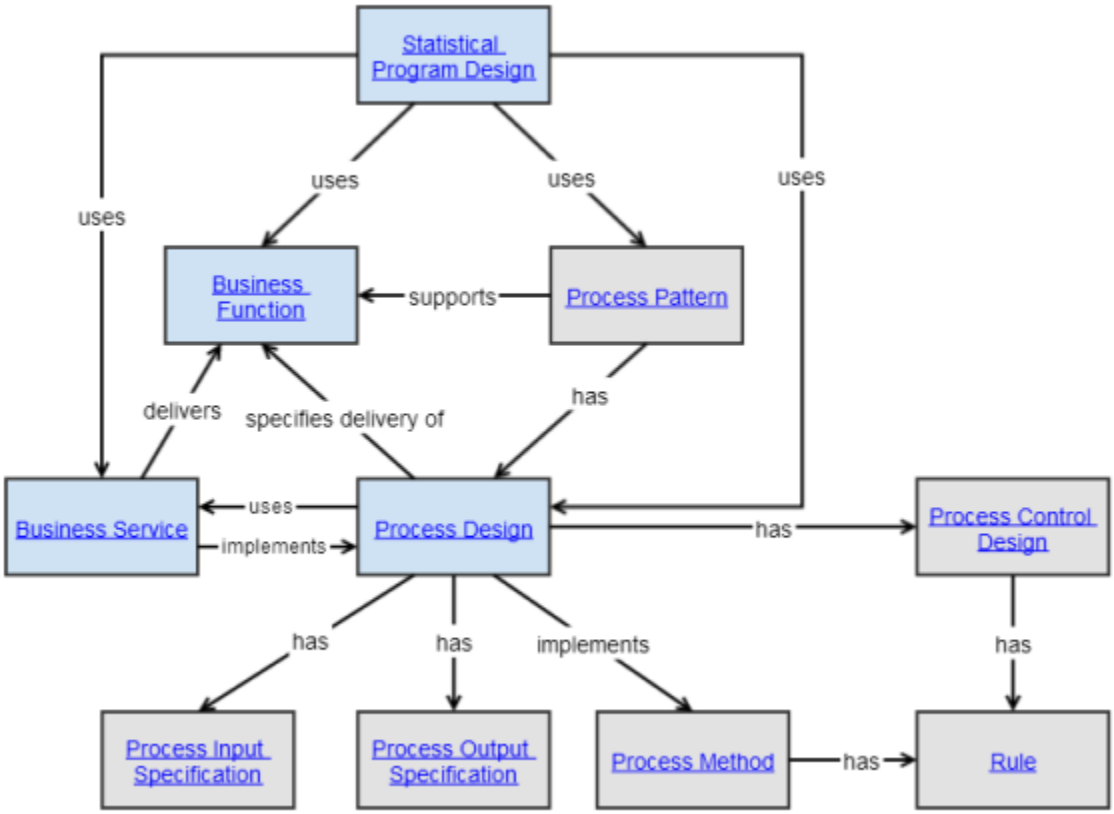

Figure 5. Use of re-usable Business Services

31. It is not always necessary for the S*tatistical Program* 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).

32. 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.

33. *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.

D: Running Processes

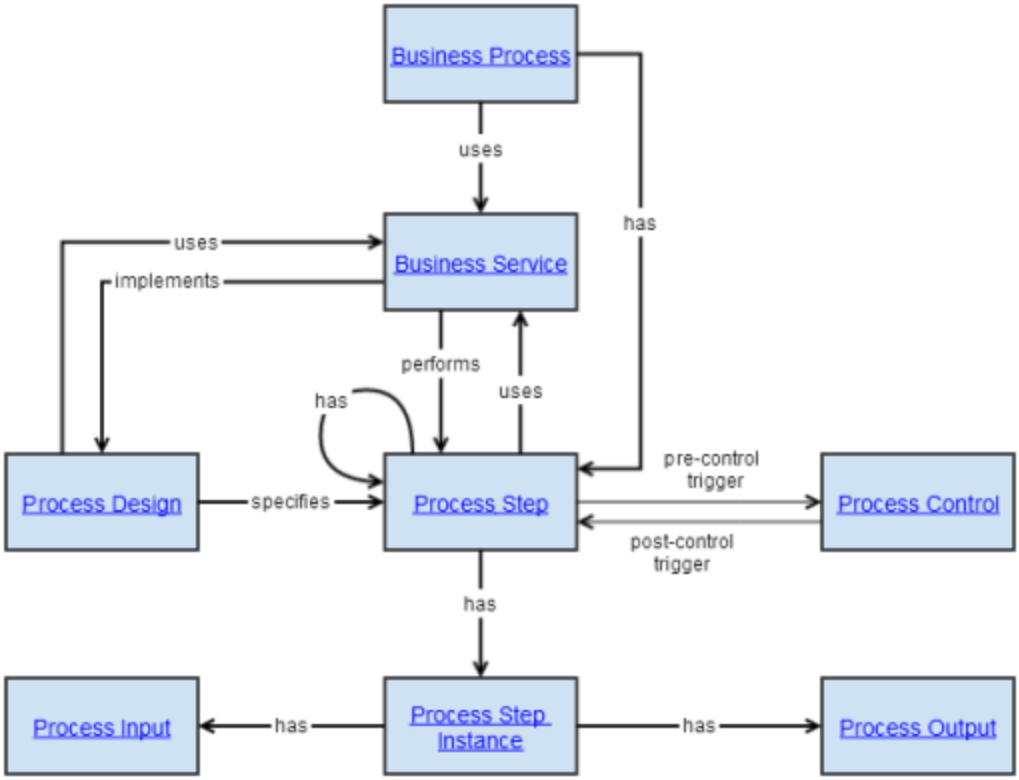

Figure 6. Run Process

34. A *Statistical Program* 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*.

35. 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*.

36. 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.

37. 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*.

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

E: Exchanging Information

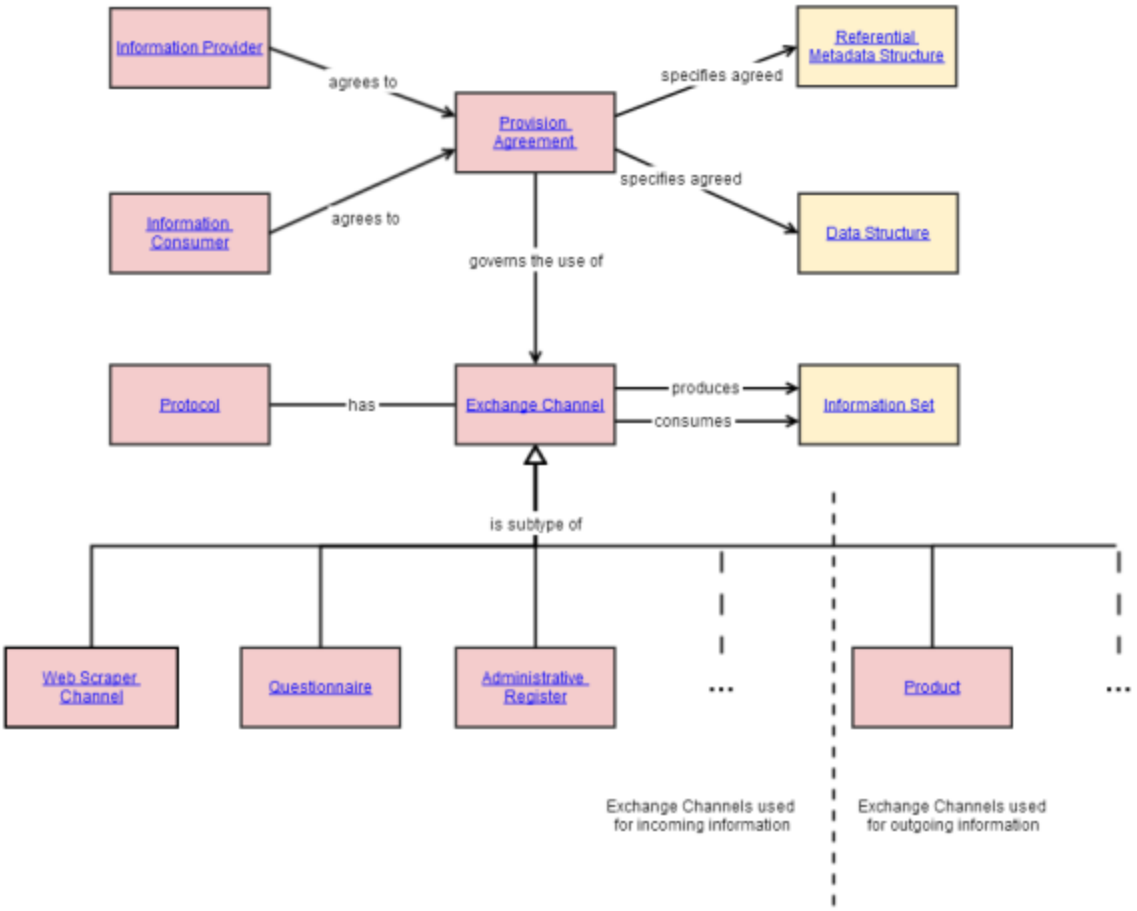

Figure 7. Exchange Channels

39. 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.

40. 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.

41. 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.

42. 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).

43. 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.

F: Collecting Information

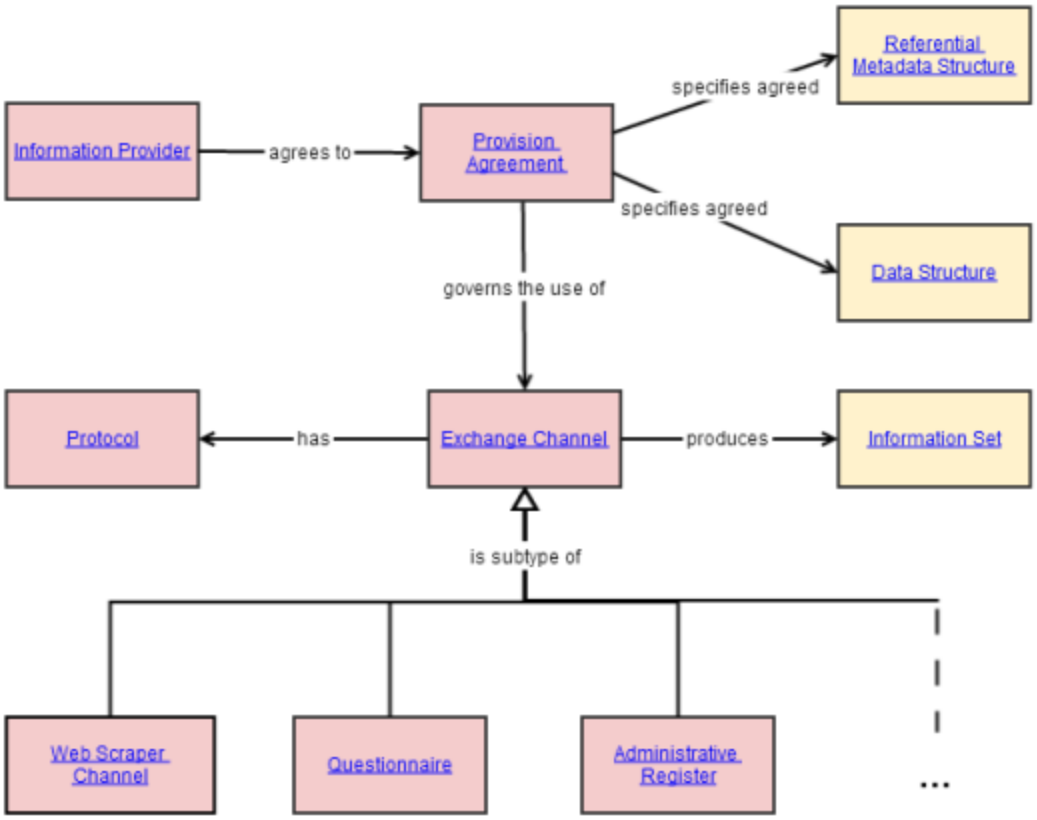

Figure 8. Exchange Channels for collecting information

44. 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.

45. 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.

G: Processing and Analyzing Information

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

47. One can understand the statistical production process from a data-centric perspective[1] . 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.

48. 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*.

49. 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.

50. 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*.

51. 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 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.

H: Disseminating Information

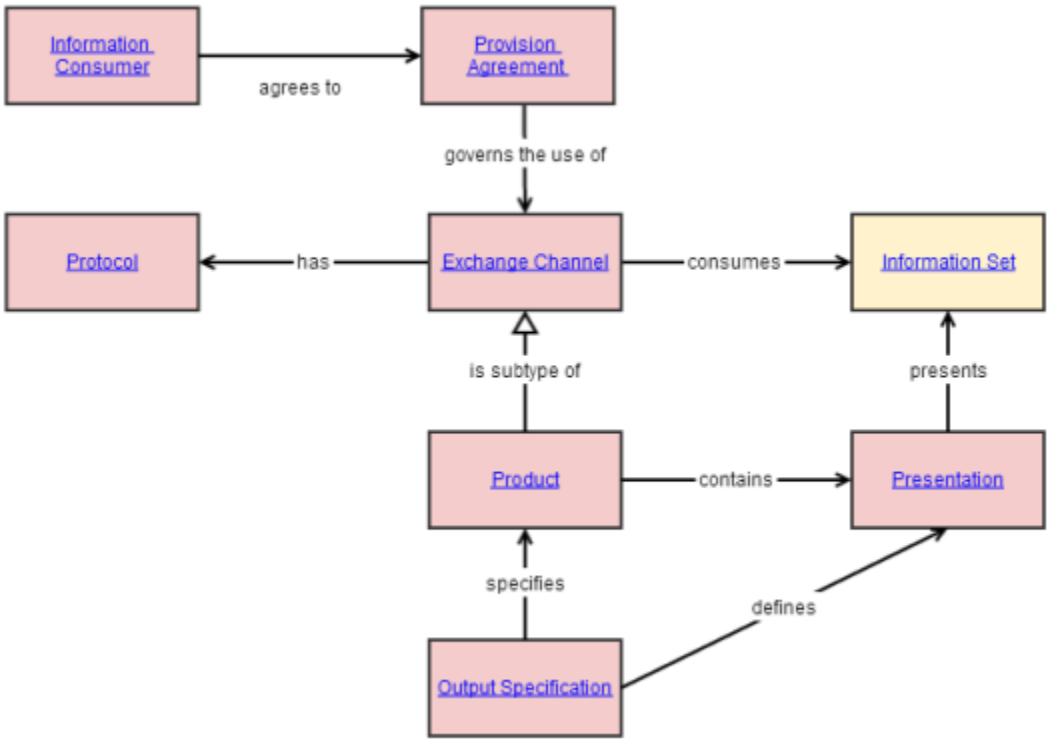

Figure 9. Exchange Channel for disseminating information

52. A statistical organization disseminates statistical information to an *Information Consumer.*

53. 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.

54. 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.

55. 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.

56. 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*.

57. 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.

58. 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.

---

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