Issue 37: CBS mapping of Acquisition Activities

GSIM – Acquisition activities

Introduction

The last couple of months, Statistics Netherlands has been working on an architectural model of the data collection domain, in terms of functions and information objects. The main purpose of this model is to be the starting point for overhauling the application landscape.

Lately, we have made an attempt to map our concepts to GSIM. We are not satisfied with the results of our attempt and are left with quite a number of questions regarding GSIM and its application in our local context.

The purpose of this document is spark discussion in the international community by explaining (in broad terms) our model of the data collection domain and showing a little of the problems that we encounter in our attempt at mapping our model to GSIM.

We believe that part of the problem is the positioning of GSIM as a vocabulary for describing statistical concepts. In the current version of GSIM, we see an unbalance in the degree of abstraction. Certain areas are very abstract, other areas in the model seem to be on a more logical or even almost physical (implementation) level. Also, some parts are very tailored towards certain applications. As our intent is to create a model that is abstract (enough) to cover a number of current and future developments, we see problems in the use of GSIM in those areas that are on a more logical level or that are focused on certain concrete applications. These concrete applications t tend to be current practices rather than future ones.

By “application” in this sense we mean making conceptual functions more specific by describing not only *what* the function does, but also *how* it does it, for instance by applying a certain (set of) method(s). Since there are usually many ways in which a conceptual function can be realized, it seems restricting to model only a few possible ways of implementing a certain concept.

SN model of data collection domain

In this section, we highlight, in our own language[1], the most important concepts out of our new architectural model. The model itself contains a complete formal description of the domain in terms of functions and information objects, expressed in Archimate. In the sequel, **bold** names indicate (candidate) information objects.

The purpose of data collection is to collect data about certain aspects of certain objects in a **Universe**. In most cases, the Universe is the real world. Objects in the real world are Persons, Businesses, Cars, but also more abstract things like movements, transactions, etc. However, a Universe may also be a representation of the real world, for instance in a registration. Even the Internet may serve as a Universe.

Usually, a Statistical Organization is interested in (the characteristics of) a related set of objects from the Universe. For instance, Household – Person – Trip – Movement. Each of these objects is defined as an object type. Each object type is related to its own **Population**. For each object type, also a list of characteristics of interest (the variables) is defined. The set of object types and their respective characteristics, we call an **Obse rvational Unit Type**. As follows from the example given, a Observational Unit Type may be a complex type, usually consisting of multiple object types. An **Observational Unit** is an instance of Observational Unit Type and is in itself a complex object.

In most cases, a Statistical Organization does not directly collect data from the Universe itself. It uses agents to observe (parts of) the Universe and to provide the observed characteristics as data to the Statistical Organization. The agent is called an **Observer**, the data provided is called an **Observation**. Traditionally, the Observer is also called a Respondent. The Observer has access to and knowledge about a certain part of the Universe.

Observer is a role that may be played by natural persons. Increasingly, however, we see automated agents being used for observing Universes. Examples of these are Internet Robots scanning web sites, magnetic loops in the pavement observing the passage of vehicles or query engines making selections from a registration.

Observational Units are often found through **Sample Frames** and **Samples.** In general, the **Sample Unit** is a part of, i.e. is one of the object types within, the Observational Unit. Usually, the Sample Unit is the top of the tree of objects to be observed by a single Observer. Often, the Observers who will be observing the Observational Units are also found through Sample Frames and Samples.

A Universe is not always “public”. That is, in many cases we need to get permission to access (parts of) the Universe. A **Responsible Unit** is (the role of) a natural person who has control over the part of the Universe identified by a Sample Unit. A Responsible Unit provides access to this part of the Universe, but in many cases also is accountable for the process of Observation, i.e. has control over the Observer. In fact, in many cases, he/she is (i.e. also plays the role of) the Observer.

The Observer must be told what to observe. For this, a Statistical Organization provides an **Instruction[2]**. We distinguish two types of Instructions. In the first place, the Observer must be told which (types of) objects in the Universe (Observational Unit) are of interest. We call this an **Identification-Instruction**. Note that through following the Identification-Instruction, the Observer identifies objects belonging to the set of Observational Units in the field rather than through samples. Secondly, the Observer must be told which characteristics of those objects to observe. This, we call an **Attribute-Instruction**. In current practice, in primary data collection, these instructions are combined and conveyed by means of a questionnaire. For an automated Observer, the instruction may take the form of a Query (SQL or web service call). By the way, a questionnaire usually is also the carrier for conveying the Observation back to the Statistical Organization.

Instructions (of both types) often need to be personalized. At least they need to be individualized for each individual Respondent. This leads to a **Personalized Instruction**.

Both Personalized Instructions and Observations are information objects. But they are also business objects in the real world. Apart from “content”, they may have multiple forms. Both content and form(s) need to be designed. The Instructions mentioned earlier being the content designs. In addition, we need an **Instruction Form Design** for each Mode.

For the purpose of communication between the Statistical Organization and the external parties (Responsible Unit and Observer), for instance in order to transfer the instruction to the Observer and get the Observation back, a communication **Connection** is needed between the National Organization and the external party. The purpose of a Connection is to bridge the distance between the Statistical Organization and the external party. “Distance” in this context means both physical distance, but also distance in semantical context, language and general posture. A Connection therefore has two (and only two) endpoints (or interfaces) and its purpose is (the effective and efficient) exchange of information between these endpoints. A Connection has a direction (one-way in either of the two directions or bi-directional).

A **Mode** is a way of communicating with a Respondent with the ultimate goal to acquire data. Mode therefore is behavior that is aimed at the effectiveness of communication. It is the way in which the exchange of information takes place. Specifically, Mode is the behavior of a Connection on the external endpoint. Mode is the way the Respondent experiences the communication. Examples of Mode are: spoken word, face to face; spoken word, telephone; screen & mouse/keyboard; paper & pencil; electronic Application to Application. On a different level, modes may be classified as person-to-person; person-to-application; application-to-application.

A Mode determines the quality of the communication exchange at the endpoint of a Connection. A Mode can be interactive or not. A Connection may support multiple Modes, and may even switch between Modes in the course of an exchange. For example, it is possible for a Channel to send an instruction or question in one Mode and receive the answer in another Mode.

A **Channel** is used to facilitate Connections for the communication between the Statistical Organization and the Observer of the Universe. A Channel establishes a Connection between the parties involved for the duration of a communication exchange. A Channel is defined as a facility to transfer information between two parties (independent of the content of that information or the reason for the exchange). A Channel has functionality and internal workings. A Channel is capable of facilitating (establishing) certain (types of) Connections.

The functionality provided by a Channel consists of two types: Transportation and Transformation. Transportation bridges the physical distance between the parties involved in a Connection. Transformation bridges the gaps in form, semantics, language, etc.

Note that the concepts of Connection, Mode and Channel are not specific for the data collection domain. Also in dissemination and in fact everywhere in the statistical value chain where data is handed over between steps and phases we see these concepts.

Mapping to GSIM and understanding GSIM as such

We have made an attempt to map our concepts to the GSIM model. Unfortunately, we did not find this straightforward. We struggle to understand the definitions and descriptions of GSIM, and were not able to make a clear mapping of our concepts.

Concept SN	**Concept GSIM**	**Comments from Thérèse Lalor (just because it was easier to associate them here than trying to jump back and forth between the comments and this text)**
Universe	There is no matching concept in GSIM. We are not sure there needs to be one.	
Population	We are not sure whether GSIM recognizes the fact that each object type in the complex Observational Unit Type has its own Population	**GSIM has an object called "Population",** **GSIM v1.1 will add an object called "Unit Type", we should check the relationship between this new object and "Population"**
Observational Unit Type		**GSIM v1.1 will add an object called "Unit Type",** **This object will not have subtypes, but an organisations could choose to implement subtypes if they wanted**
Observational Unit		**GSIM has an object called "unit".** **In GSIM v1.0, there was a subtype called "Observation Unit" but this will be removed in v1.1, allowing organisations to implement subtypes that are meaningful to them.**

Observer	In our opinion, the matching concept in GSIM should be Data Provider. But the definition in GSIM does not allow this to be an automated agent. Table 1 in paragraph 27 seems to indicate that a clear definition of this concept in GSIM would help clarify things, since it seems that GSIM considers an automated agent to be an Instrument Implementation. This is confusing, since it mixes the provider side and the consumer side of the relation.	
Observation	The structure of an Observation is a (Unit)DataStructure. The Observation itself is abstract, it may take the form of a DataSet, but may also exist in the mind of the Observer and be transmitted as sound.	**I suspect that this is the GSIM object "Datum" (glossary says "A datum is the actual instance on data that was collected")** **It could also be "Data Point"?** **transmitted as sound = "Non Structured Dataset"**
Sample Frame		**Samples and Frames are not in GSIM. I think in a discussion a few weeks ago where we decided that this was not going to make it into v1.1.** **If I stretch my brain all the way back to v0.3, didn't we say that frames are a type of dataset (composite dataset i think)**
Sample		**See above**
Sample Unit		**GSIM has an object called "unit".** **In GSIM v1.1, there will be no subtypes allowing organisations to implement subtypes that are meaningful to them.**
Responsible Unit	We did not find a direct match for this concept in GSIM. It might be related to Data Provider.	**GSIM has an object called "unit".** **In GSIM v1.1, there will be no subtypes allowing organisations to implement subtypes that are meaningful to them.**
Instruction	There is no matching concept in GSIM. Instrument is not the same, an Instrument is the carrier of Instructions. The concept of Instruction within the data collection domain has a conceptual counterpart in the dissemination domain in the Output Specification: the same concept, different application.	**This way of thinking about data acquisition is a bit foreign to me, however, I don't that that something is missing from GSIM, rather that we may need to apply some GSIM objects in a way we have not yet considered?**
Identification Instruction	GSIM does not make the distinction between Identification of relevant objects in the Universe and the observation of characteristics of such identified objects.	**This way of thinking about data acquisition is unfamiliar to me, however, I don't that that something is missing from GSIM, rather that we may need to apply some GSIM objects in a way we have not yet considered?** **Does anyone in GIG have a similar approach?**
Attribute Instruction	GSIM does not make the distinction between Identification of relevant objects in the Universe and the observation of characteristics of such identified objects.	**This way of thinking about data acquisition is unfamiliar to me, however, I don't that that something is missing from GSIM, rather that we may need to apply some GSIM objects in a way we have not yet considered?** **Does anyone in GIG have a similar approach?**

Personalized Instruction		**This way of thinking about data acquisition is unfamiliar to me, however, I don't that that something is missing from GSIM, rather that we may need to apply some GSIM objects in a way we have not yet considered?** **Does anyone in GIG have a similar approach?**
Instruction Form Design		**Is this something near the Acquisition Design objects?**
Connection	No match in GSIM	**This makes me think of CORE (probably not surprisingly!),** **It seems a bit related to "Data Channel"?**
Mode	Mode as a concept is related to Connection (and therefore to Channel), but it depends on where the concept is applied, on which end of the Connection the Mode is important enough to be modeled. In the data collection domain, the Mode is on the external end, i.e. the “input” end of the Connection in terms of the flow of statistical data. In the dissemination domain, the Mode is more important on the “output” end of the Connection, i.e. the end of the external user of statistical products. The examples given in Table 1 (paragraph 27) in our opinion should be revisited. The introduction of Instrument Implementation unnecessarily complicates the matter. We wonder whether a Set of Requirements as Instrument is correct, as this would indicate Design Phase activities?	**That table in paragraph 27 really needs some work!**
Channel	Channel is a generic concept and has its applications in data collection as well as in dissemination. In fact, we believe that even in Production there may be a need to apply the concept. More specifically, we believe that any exchange between Services in CSPA in fact means applying the concept of Channel. The concept as currently defined in GSIM is different from ours, and does not clearly define both endpoints.	**We should think about this.**

While studying GSIM, we found some other areas where we do not clearly understand the definitions and descriptions. We list just a few of them here.

Instrument and Instrument Implementation are GSIM objects that seem to be more concrete in nature than other objects in GSIM. We question whether (in the current state of affairs) detailed modeling of such concepts is needed in GSIM.

We do not understand the true meaning of Channel Activity Specification and Channel Design Specification. As specified, we do not see them as being specializations of Data Channel.

The relationship between Channel and Data Resource is unclear, there seems to be inconsistency in GSIM with respect to what a Data Resource is.

Conclusion

In our exercise, we encountered a number of problems in truly understanding GSIM concepts, definitions and explanatory texts. We believe this is partly due to the language used, but a large part of it is due to inconsistencies and the fact that GSIM is a mixture of objects on different levels of abstraction. We find it hard to understand GSIM in our own context.

We believe that certain areas of GSIM are on a level of abstraction that is too low and therefore are restricting in modeling future applications, modes of operation.

[1] We apologize if some of the English terms used in this document as a translation of our Dutch terms are unclear or confusing. In future, we hope to rectify this when we succeed in mapping them to GSIM terminology.
[2] In fact, **Request** might be a better term, as we are requesting an Observer to provide certain information. Note that Instruction here refers to a generic design, valid for all Observers in the survey.