Distr. GENERAL

01 May 2013

WP.3

ENGLISH ONLY

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE
EUROPEAN UNION (EUROSTAT)

ORGANISATION FOR ECONOMIC COOPERATION
AND DEVELOPMENT (OECD)
STATISTICS DIRECTORATE

Work Session on Statistical Metadata
(Geneva, Switzerland, 6-8 May 2013)
Topic (i): Metadata standards and models

GSIM STANDARDS AND MAPPING: PRELIMINARY FINDINGS AND STATUS

Working Paper

Prepared by Arofan Gregory, Metadata Technology North America

I. Background and current status

1. One of key priorities for work related to statistical modernisation in 2013 is to support the enhancement and implementation of the standards needed for the modernisation of statistical production and services. As such, at the start of 2013, a project called Standards and Frameworks for Statistical Modernisation was initiated. This project is important for the official statistics community because a lot of effort has gone into the development of standards such as the GSBPM and the GSIM. To realise the full benefits of this investment, it is necessary to ensure continued coordination during the implementation phase, and to ensure that the lessons learned during implementation are shared, and reflected in enhanced versions of the standards and their associated documentation.

2. Under this project, several task teams were formed to carry forward the work. One of these groups is focused on creating agreed mappings to two important implementation standards - SDMX and DDI. The work of this group is still in progress, but it is not too early to describe the current state of progress, and to look at the group's preliminary findings.

3. In order to be clear about how GSIM stands in relationship to other standards, it is useful to have terms for the different types of models which we are describing. GSIM itself is often described as a "reference model" and as a "conceptual model" - it is both of these. The idea is that GSIM is not designed to be directly implemented, but is instead useful for describing implementations done with mode specific models. These specific models - "implementation models" - are at a more detailed level, and both SDMX and DDI fit into this area. Thus, when we map GSIM to SDMX and DDI, we are mapping a conceptual model to an implementation model. There is a more specific level of modelling, too, which we will describe as an "application model" - this is the model which exists within any specific software application (and so is not necessarily a harmonized “view”), and is itself a mapping from the implementation model to the actual software itself. Often, but not always, these last two levels of model are very similar.

4. The GSIM mapping group is a very diverse group: it includes members from national statistical organizations and central banks, the standards bodies, and international organizations. The group holds bi-weekly teleconferences. At the initial meetings, the overall approach was discussed, as were the goals of the effort. It was recognized that without a body of implementation experience to work from, the initial output from the group might not be the final word, but would instead be something which would need to be validated and adjusted based on implementation of GSIM using the standards. One significant agreement was that any given GSIM object need not be mappable to both standards - so long as an implementation of some portion of GSIM was supported by one standard or the other, that was acceptable. In some cases, standards other than DDI or SDMX could be used as well, depending on what makes the most practical sense for implementers.

5. The group agreed to divide into two sub-groups, one focused on the GSIM-DDI mapping, and the other on the GSIM-SDMX mapping. There is some overlap in membership between the two groups, so that a degree of coordination is achieved between the two groups.

6. The overall approach was this: a first pass would be made based on the initial mapping work which was provided in the GSIM 1.0 package, for both SDMX and DDI, and specific areas would be identified for more detailed work or specific actions. In some cases, the more detailed work might be done by a task-team consisting of a few members of the sub-group, whose work would then be reviewed by the sub-group as a whole. One task-team was established to look at key administrative metadata, and this sub-group involves members from both the SDMX and the DDI sub-groups.

7. Currently, the first high-level discussion of the existing mappings has been completed by both the DDI and SDMX sub-groups, and gaps and areas for further work have been identified. As this work progresses, there will be a more focused look at specific areas of the mappings, informed by the use cases identified by the HLG Common Statistical Production Architecture project and by other GSIM implementations currently on-going at the institutions of some participants, such as Statistics New Zealand and the Australian Bureau of Statistics.

II. The goals of the mapping work

8. GSIM 1.0 was produced on an aggressive timeline, using a methodology based on the popular "agile" methodology for software development, using a series of "sprints". It was recognized that the first version of GSIM would require revision, based on initial implementations. The standards bodies which produce SDMX and DDI also recognized that GSIM is important to them for their future development and adoption. Further, the SDMX-DDI Dialogue facilitated by UNECE had early on recognized that GSIM would likely form the basis on which the two standards could be aligned.

9. Thus, a number of possible goals were identified for the work of the GSIM standards-mapping effort.

10. The most obvious of these is the mapping itself, showing how the objects in GSIM could be implemented using SDMX and DDI. It should be noted that in some cases this mapping is a clear one-to-one mapping; in other cases, there may be a difference between the granularity or scope of the GSIM object and the corresponding construct in the target standard. In these cases, some condition or qualification must be specified to provide a useful correlation between GSIM and the implementation standard. In some cases, a mapping of this type is possible, but might be of questionable utility, and these cases have been identified in the spreadsheets used to capture information about the mappings. Such cases are the focus point for further exploration as the work proceeds.

11. It should be noted that the GSIM-DDI mapping is working with three versions of the DDI standard: the latest version of the DDI-Codebook line, version 2.5; the current production version of DDI-Lifecycle, version 3.1; and the next version of DDI-Lifecycle, now being finalized after public review, version 3.2. For the GSIM-SDMX mapping, both version 2.0 and 2.1 are being considered, although these versions have been found to be fundamentally similar for the purposes of the mapping work thus far.

12. Another major output of the GSIM mapping work is to produce feedback to the groups developing and maintaining SDMX and DDI. Because there has been a commitment to work collaboratively, it has been easy to communicate with these groups – the DDI Alliance is already collaborating, albeit in an informal fashion, and it is anticipated that a similar collaboration can be established with SDMX. Specific actions have been agreed within the GSIM mapping task teams to formulate formal input to the DDI Alliance and the SDMX Initiative, but even without formalized input there are already some changes being considered to the final DDI version 3.2 which will make it a better vehicle for GSIM implementation, notably in the area of variable description for microdata. This is promising, as both SDMX and DDI provide coverage for some areas of GSIM, but not necessarily in the best possible way. Working together, GSIM, SDMX, and DDI could provide a better path to implementation in future.

13. Finally, some of the mapping work has highlighted issues which will be forwarded to the GSIM Implementation Group, for potential inclusion in future revisions of GSIM itself. Thus far, there have not been many issues identified, but it does appear that some portions of GSIM could use more detail within the GSIM model itself, to better facilitate implementation. This is not a surprising discovery, but it is useful to have the input for future work on the GSIM model.

14. It is important to point out that although the mapping work has identified some specific gaps in coverage for the entirety of the GSIM model, the approach to the mapping work has been quite conservative: it is accepted that the whole of GSIM will need to be implementable in some fashion, but that this does not justify trying to force SDMX and DDI to perform tasks for which they are not designed or well-suited. Identifying the gaps is a process of investigation, and the existence of a gap is itself useful information resulting from the mapping work. The group is not intending to produce new standards or solutions for these gaps - that has been judged to be outside the scope of this effort, although it will certainly be a topic of discussion among the implementers of GSIM.

III. Coverage of GSIM by SDMX and DDI

15. Given the current state of the work, it is possible to give a general sense of where there is good support for GSIM, where there are gaps, and where further action is needed. We will look at each area of the GSIM model, the Business, Concepts, Production, and Structures areas, and look at where each standard provides support or fails to do so, and what actions are seen as necessary moving ahead.

A. Business group

16. This part of the GSIM model covers a range of objects, some of which are very well-supported by the standards, and others not so well, or indeed not at all. One set of objects is concerned with statistical activities and design. This area is not very well supported by DDI at the moment, although it is felt that given DDI's "lifecycle" design, the standard could support this portion of GSIM. In the next major revision of DDI (the version to be produced following version 3.2) it is anticipated that there will be a "model-agnostic" mechanism for describing business processes and provenance. It is felt that such a model should support this portion of GSIM, and that the gaps here should be formulated as input to the DDI revision process.

17. While SDMX has a model for documenting processes, this is very generic, and may or may not provide a suitable implementation model. This would depend on how other aspects of the implementation were being conducted, as the SDMX process model can usefully reference objects which are expressed in SDMX, but is not as useful for referencing non-SDMX objects.

18. Another part of the Business Group in GSIM describes instruments for data collection. In general, this portion of the model is very well supported by DDI, although some questions for clarification remain regarding instrument control. This is an area for further work within the mapping team. SDMX provides no specific mechanism for describing survey data collection, or collection activities

per se (although this can be done using the Reference Metadata features of SDMX). Thus, this was seen as a portion of GSIM most appropriately supported by DDI.

19. Another portion of the Business Group in GSIM concerns specification of needs and the business case for statistical activities. While there are some parts of DDI which support some objects, in general DDI is very weak here. This is especially true as it is felt in future that these GSIM objects should support automation of processes, and the support provided by DDI is simply documentary in the current versions. It was decided that requirements for future inclusion in DDI should be formulated.

20. Whilst SDMX was generally lacking in explicit support in this area, much of the metadata that passes between the processes supporting the needs and business case can be expressed using the Reference Metadata features of SDMX). Again it was felt that the design of DDI was better suited for supporting this portion of the model, if changes could be made to that standard.

21. There are also a set of objects related to the GSIM Data Channel, and in the case of both SDMX and DDI it is felt that there is potentially some support for this (a Data Channel provides data sets which could be described using SDMX and DDI) but that there was no obvious mapping here. This area needs further discussion, to identify what changes should be recommended to provide implementation support for GSIM.

B. Concepts group

22. The Concepts Group of GSIM contains a set of objects which could be described as the "classic" metadata which we manage within our statistical systems: classifications, variables, concepts, populations, statistical units, etc. In the case of both SDMX and DDI, this can be problematic, because, while GSIM is designed to support such activities as classification management, neither SDMX nor DDI are: they both describe the uses of classifications, but were not intended to support classification management itself.

23. In many cases, both DDI and SDMX provide support for the objects in this part of GSIM: Concepts, for example, are supported by both standards, as are codelists. However, the uses of concepts within GSIM are broader than in either SDMX or DDI, something which provides a lot of power within the GSIM model. However, the way in which concepts are modelled in GSIM, in relation to other core metadata, is being considered for inclusion into version 3.2 of DDI-Lifecycle. The way in which GSIM models variables, too, will be implemented in this version of DDI as well, even though there is fairly good support for GSIM variables already in DDI-Lifecycle 3.1. Variables are an area where SDMX has "collapsed" the metadata structures - a variable in GSIM as a maintained and reusable object is most easily mapped to a Concept in SDMX as the Concept in SDMX can have a “default representation” but from an implementation point of view this is not a very satisfactory mapping.

24. It should be noted that this part of the GSIM model is full of abstract classes - classes which are never intended for direct implementation (although you might in fact implement them in an application model if they make sense). These mappings are considered secondary within the GSIM mapping group, as an implementation would never directly instantiate an abstract class (which is the reason it is abstract.)

25. Another set of the objects in this part of the GSIM model are those relating to populations and statistical units. There is no explicit support for these constructs in SDMX. DDI does provide support for many of them, however, although again GSIM provides a richer model in some ways.

26. It is important to note that for this part of the GSIM model, it is very possible that DDI will be modelled directly after what is found in GSIM, although with a stronger implementation focus. This is currently being discussed within the DDI community. Now that more statistical agencies are starting to adopt DDI, it is seen as appropriate that DDI supports classification management, and also that the power of the GSIM model would provide a good basis for this portion of the DDI implementation standard as well. Thus, in future, we will probably see a very strong alignment

between GSIM and DDI. For SDMX, there are no plans to support this part of the model beyond the current support for classifications.

C. Production group

27. The Production Group in GSIM is very much centred around “actionable” metadata. The support for actionable metadata is not very good from either SDMX or DDI. While SDMX does provide a model of processes which maps quite well to equivalent objects in GSIM, it is more useful for documenting business processes than it is for managing statistical processing (to be fair, SDMX was not designed for this purpose). Having said this there is an active group in SDMX which is developing a standard way of formulating “expressions and calculations” that can be used to specify the operations at a granular level, such that a program can “read” the metadata and compose the expression required in whatever computer language is appropriate. It is probable that this will become a part of a future version of SDMX. While DDI does contain some support for some of the objects in this group, GSIM contains a much higher degree of detail. Thus, it is seen as a useful action to submit the gaps as potential requirements for inclusion in future versions of DDI. It should be noted that DDI-Lifecycle version 3.2 provides better support in this group than DDI-Lifecycle 3.1, but even so there are some important gaps here.

28. The lack of support from DDI and SDMX here is not seen as a critical lack, because there are many other standards which could - and are - used in this area, such as BPMN, BPEL, and others. For this reason the key for standards mapping is to consider which standards provide the best support for GSIM, and then to determine how these standards might be implemented in a coordinated fashion with SDMX and DDI. This is clearly an area that needs to be discussed not only within the GSIM mapping team, but within the GSIM implementation community more broadly.

D. Structures group

29. This is the part of GSIM which aligns most closely with the implementation standards: both SDMX and DDI provide excellent support for many of the objects here, as both were designed for describing the structural metadata used by data sets. GSIM provides two possible ways of describing the structure of data: as unit-record data, or as dimensionalised data. In the case of SDMX, it is designed to use a dimensionalised approach to data description, but it maps very cleanly onto this portion of the GSIM model. Further, many of the constructs found here such as provision agreements and data flows are constructs which are also found in SDMX.

30. DDI can be used to describe data either as unit-record data or as dimensionalised data, and so provides support for both approaches. It should be noted that DDI was explicitly designed to align with the SDMX view of dimensionalised data structures, so strong support in this area from both standards is not surprising. Support for such objects as data flows and provision agreements does not exist in DDI, however. For implementation, SDMX probably provides better support here, but for specific uses both standards are quite strong.

31. It should be noted that one portion of this group in GSIM is not well supported by either standard, although again SDMX is probably better than DDI: that is in describing dissemination activities. Further investigation is needed into how best to support this portion of GSIM, with SDMX being a more likely candidate to receive requirements here. Specific requirements still need to be identified and discussed, however.

IV. Conclusions

32. The GSIM mapping work is already producing some interesting results, and promises to become even more interesting as the more detailed mapping work continues. It will provide not only a useful tool for implementers of GSIM, but also will serve to influence the further development of DDI and SDMX. The work is only an initial effort at determining the best way to implement GSIM.

33. It should be remembered that GSIM, as a conceptual model, does not place demands on implementers such that they change their production systems to support everything it contains. Rather, the production of statistics should be supported by whatever portions of GSIM are needed. Thus, while there is not complete support for GSIM within the SDMX and DDI standards, many very critical portions of GSIM are already supported, and will be better supported in future. Thus, we can see GSIM as a useful tool for driving convergence within the domain of statistical metadata, and one which will be to everyone's benefit in the longer term.