Distr. GENERAL
26 April 2013
WP.24
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 (iii): Metadata in the statistical business process

PILOTING GSIM AND GSBPM AT THE INTERNATIONAL MONETARY FUND

Working Paper

Prepared by Gareth McGuinness, Michaela Denk and Alberto Sanchez, IMF[1]

I. Introduction

1. The IMF Statistics Department plans to use the Generic Statistical Information Model (GSIM) and the Generic Statistical Business Process Model (GSBPM) when upgrading or re-engineering existing processes as part of its on-going "Streamline, Standardize, Automate" program.[2] These opportunities will be used to build a holistic view of the end-to-end statistical process.

2. The Department piloted the effort by identifying the sub-processes and information objects used in its internal Integrated Monetary Database (IMD) product.

3. Part II of the paper outlines the process we used for the pilot project. Part III reports the findings of the pilot, and how we plan to use its results to improve the performance and efficiency of the IMF's work to collect, process, and disseminate data. Part IV examines our experience in the pilot, and where GSIM and GSBPM can be useful for statistical offices, and suggests possible ways to improve the models (particularly GSIM).

1 The views expressed herein are those of the authors and should not be attributed to the IMF, its Executive Board, or its management.

2 Full documentation on GSIM and GSBPM can be located on the UNECE website at http://www1.unece.org/stat/platform/display/metis/METIS-wiki.

II. Our pilot's process

4. We followed a three-step approach to map the IMD statistical process to GSIM objects and GSBPM processes. We gathered information, mapped to GSBPM and GSIM, and integrated processes and objects. This part of the paper explains what we did in each phase of the work.

A. Information gathering

5. We started by meeting IMD data owners and project managers to learn their process. We sought their views on our proposed approach while sharing our thoughts about its value.

6. After some discussion, we identified that it would be most efficient to present a questionnaire to the project managers to elicit a structured response. We based the questionnaire on GSBPM sub-processes, adapted to match key IMD project features.

7. We received a very detailed and quick response. For example, one question asked for a sketch of how the new IMD process would tackle former process inefficiencies; we received back a diagram (Figure 1), plus a detailed description of each step.

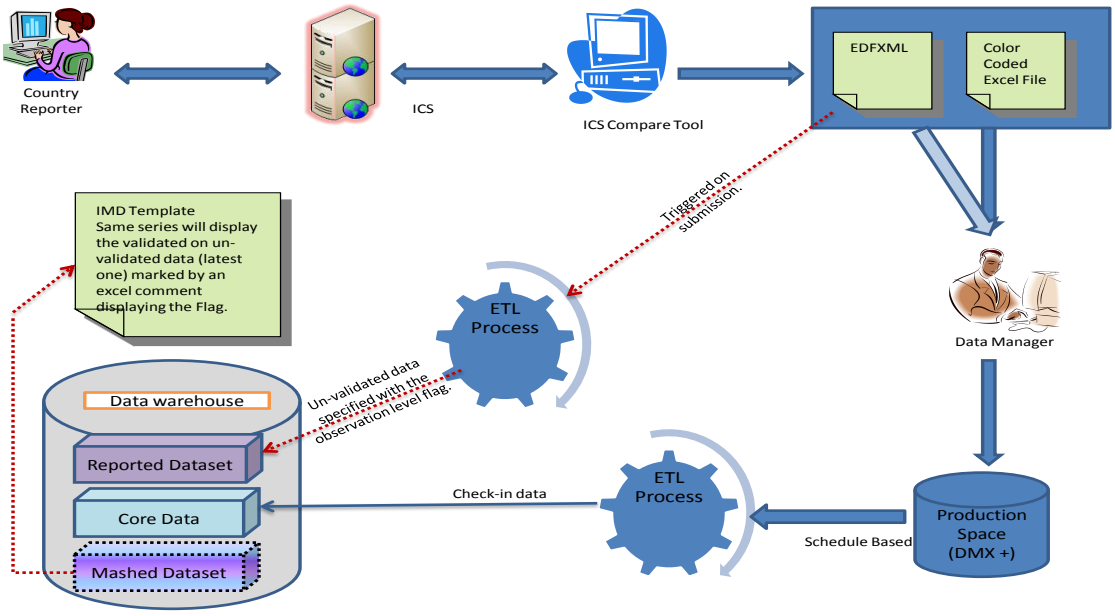

Figure 1 – New IMD process.

8. The final step consisted of clarifying unclear responses and identifying unanswered questions.

B. Mapping to GSBPM and GSIM

9. We met once a week, for less than two hours. We extracted answers from the completed questionnaire and classified items into three categories: processes, objects and other.

10. We then started the mapping process. Each member of the pilot group mapped a part of the whole and we discussed our findings in our meetings.

11. We examined GSBPM sub-processes first and then GSIM objects. The latter took the most time, as we required diagrams to show relations between objects and found it difficult to map IMD objects to GSIM. To manage this, we considered objects in the GSIM Production Group later, when we matched sub-processes with inputs and outputs.

C. Integration between processes and objects

12. Finally, we focused on the GSIM Production Group and identified objects forming GSBPM sub-process inputs and outputs. Achieving this allowed us to complete the picture of how information flows through the process and how input objects transform into output objects.

13. Matching objects to inputs and outputs took some time, but proved very useful in raising questions and stimulating ideas, which helped us to identify possible missing objects and processes in both the IMD process and in the generic models. We discuss these missing objects and processes in the next two parts of the paper.

III. Findings related to analysis of the IMD process

A. Process view (GSBPM)

14. The IMD is an existing product. Several months before we started our GSIM/GSBPM exercise, the IMD team had analyzed the existing IMD process and developed a to-be process to make new IMD data available to users in a timelier manner.

15. Our analysis took into account the existing and to-be production processes, but not the project to review and evaluate the existing process, and design and build the to-be process. We focused on GSBPM processes 4 (Collect) to 7 (Disseminate), considering processes 1 (Specify Needs) to 3 (Build), 8 (Archive) and 9 (Evaluate) out of scope. They had either taken place when setting up the existing IMD process or during the Straight-through IMD project (e.g. identifying the need for timelier data, re-designing and building the new process). The results of our questionnaire reflected this, primarily covering the regular production process (existing and to-be) and only mentioning a few sub-processes of GSBPM process 3 (Build).

16. In general, IMD sub-processes matched GSBPM sub-processes very well. We found it straightforward to specify the mapping. We noticed some sub-processes are more granular in the GSBPM than in the IMD process and vice versa. The IMD ETL (Extract-Transform-Load) process provides an example, being broken down further in the GSBPM into at least 6.5 (Finalize outputs) and 7.1 (Update output system).

17. Decomposing these processes helped us identify additional objects not mentioned in questionnaire answers, as they were considered intermediary objects only within the ETL process, not inputs or outputs of a sub-process. In contrast, the IMD process sometimes proved more specific, for instance, concerning GSBPM sub-process 4.3 (Run collection), which mapped to multiple IMD sub-processes.

18. Overall we think the IMD process contains all the relevant GSBPM sub-processes; there are no apparent gaps compared to the GSBPM in terms of relevant sub-processes not being part of the IMD process. However, the level of detail of our process understanding varied and some of the documentation only exists as a response to the questionnaire we provided.

19. We conclude that following GSBPM would significantly improve our process understanding. Using GSBPM would also help us standardize documentation across products and make it easier to compare processes, thereby increasing transparency and reusability of sub-processes.

B. Object view (GSIM)

20. We found mapping IMD objects to GSIM trickier than mapping sub-processes to GSBPM. Strikingly, we only used 10 of the 150 GSIM objects to represent IMD objects, mainly from the Structures group. We used Dataset and Instrument Implementation (from the Business group) most frequently.

21. We identified four reasons for using so few objects.

22. First, we did not include objects relevant to the out of scope processes 1-3 and 8-9.

23. Secondly, objects from the GSIM Production group were largely missing in the questionnaire response we received. The available material showed what sub-processes take place, but did not include detailed descriptions of underlying process steps, controls and rules. We assume, though, that for each GSBPM process defined we should specify GSIM objects from the Production group.

24. Thirdly, in cases when multiple objects may match an IMD object, we picked the most generic one. The distinctions made in GSIM may become relevant when implementing a process, such as being able to differentiate “objects” at various levels (e.g. a variable) or in different roles. For our purposes mapping at the higher level made more sense.

25. Lastly, some IMD objects fell outside the scope of GSIM, being generic rather than statistical (e.g. project time frame, user training).

26. The IMD observation flag “unvalidated data” provides a good example of an “object” changing its role, i.e. occurring as various object types.

27. As long as we separate conventional IMD data and the new straight-through IMD data into separate datasets, the flag operates not as an object, but rather as an attribute of the dataset, e.g. name and/or location. In a combined dataset with two values, one validated and one unvalidated, the flag operates as an Identifier Component. In case of a mashed dataset with only the most recent value presented for each data point (either the validated or unvalidated value), the flag operates as an Attribute Component.

28. While GSIM allows these fine-grained distinctions, are they necessary? Which of the GSIM objectives does this satisfy?

29. When we integrated the process mapping and the object mapping, we were able to identify objects missing in the IMD questionnaire response.

30. We found sub-processes missing any Process Metrics. We identified where the context required a Rule, but no Rule existed. We saw Datasets appear fleetingly as intermediary objects (being intermediary output and input of one IMD sub-process spanning multiple GSBPM sub-processes). For another IMD sub-process, outreach (GSBPM sub-process 7.4 Promote dissemination products), the output objects have not yet been defined by the IMD project team, and GSIM appears to have no equivalent.

31. We also observed the opposite case: some objects described in the IMD documentation had no process using or producing them, even though we succeeded in mapping them to GSIM. These objects were outputs of the Straight-through IMD project itself, so we did not capture them by looking at the production process. GSBPM and GSIM help understand the difference between conceptualization/design processes and “run-time” processes (collection to dissemination).

IV. Using and improving the generic models

32. We found using the generic process and information models in our pilot exercise instructive. It gave us great insight into using the two models to analyze real world statistical processes. This part of the paper shows how we found the models useful and proposes improvements to the models (particularly GSIM).

A. GSBPM

33. GSBPM is a mature model. Sub-processes in the model are self-explanatory and easy to distinguish from other sub-processes. Where we could not quickly determine the most relevant sub-process, the textual description of each sub-process helped us make a decision. We feel GSBPM has the right level of detail to create shared understanding between key players.

34. We identified one area where GSBPM has a gap. The model has no dissemination services sub-process. GSIM describes (accurately, we believe) objects for such services, but no existing GSBPM sub-process has these objects as inputs and outputs. Despite this gap, we found GSBPM very usable and fit for our purposes.

35. We also uncovered a grouping of the Level 1 processes GSBPM users will understand, but which does not form part of the model. GSBPM processes 4 through 7 (Collect, Process, Analyze, Disseminate) form what we might call the "work" process of the statistical office. Processes 1 through 3 and 9, in the order 9, 1, 2, and 3 (Evaluate, Need, Design, Build) form the "change" process. Process 8 (Archive) posseses sub-process with characteristics of both work and change.

36. GSBPM contains a convenient fiction that the evaluate process happens following each instance of a statistical process. The model paper hints at the reality, where "...in some cases, particularly for regular and well established statistical business processes, evaluation may not be formally carried out for each iteration. In such cases, this phase can be seen as providing the decision as to whether the next iteration should start from phase 1 (Specify needs) or from some later phase (often phase 4 (Collect))."[3]

37. We think it more useful to consider the work process and change process as being fundamentally different. The work process produces statistics, directly creating value for society. The change process improves or degrades work processes, indirectly creating or destroying value by changing the performance and cost of the work process (or of the stock of work processes if adding or removing a work process). We realize value from the change process only when work processes improve.

38. Step one for the industrialization of official statistics is to use the change process to automate the work process. Our view is that it would be useful for the GSBPM model to explicitly support this.

B. GSIM

39. GSIM is a very new model. Version 1.0 followed intensive development involving experts from a range of agencies (including the IMF). The group of people

[3] Generic Statistical Business Process Model, Version 4.0, April 2009, paragraph 39: http://www1.unece.org/stat/platform/display/metis/Generic+Statistical+Business+Process+Model+Paper.

working on GSIM provided a comprehensive and exhaustive model, impressive given the small amount of time and high expectations.

40. The work was a first step toward achieving the initiative's six stated objectives, which we restate here with headings we have created. We will use these headings to refer to each objective.

1. **Communicate:** Improve communication between different disciplines involved in statistical production, within and between statistical organizations; and between users and producers of official statistics.
2. **Collaborate:** Generate economies of scale by enabling greater collaboration within and between organizations, especially through reuse of information, methods or technology.
3. **Automate:** Enable greater automation of the statistical production process, thus increasing efficiency and reducing costs.
4. **Innovate:** Provide a basis for flexibility and innovation, including support for the easy deployment of new statistical products and the adoption of new types of statistical data sources.
5. **Educate:** Build staff capability by using GSIM as a teaching aid that provides a simple, easy to understand view of complex information, with clear definitions.
6. **Validate:** Validate existing information systems and compare with best practice in other organizations.

41. At a high level, we sought to use GSBPM and GSIM to validate our proposed Straight-through IMD product. We wanted to understand if the processes and information objects used were part of the common set used by statistical offices and to identify possible sub-processes and objects not included in the IMD process that, if used, would generate a better product. Our project falls under objective six of GSIM – validate.

42. As we sought to apply GSIM, we used the six objectives as a reference point when we faced challenges, to help us understand if our efforts fell outside the objectives of the model. When we felt we were within the model's objectives, we sought to understand where GSIM failed to support our efforts and what changes to GSIM would help. We found significant challenges, not surprising given that GSIM is new.

43. The most common challenge we faced during our pilot project was that we had to choose from numerous GSIM objects for many of our IMD process information objects. First, we often had to decide if it made sense to represent an object as a design object, a production object or a production instance object. Then, we may have had to select a sub-type for a particular object.

44. We had the good-fortune of having a group who each had significant official statistics experience, extensive metadata skill-sets, and involvement in the development of GSIM. While we succeeded in identifying GSIM objects for each information object in our process, it left us wondering if 1) a less-suited group would have succeeded, 2) the

exercise was worth the time taken and 3) GSIM is scalable. We had negative impressions on all three counts.

45. Being invested in the idea that GSIM is an important development in official statistics, essential to realizing a vision of industrialized official statistics, we sought to understand why we felt so negative and what path will make GSIM more uplifting, thereby maximizing the chance it will be used.

C. Generic Model Value Pyramid

46. Perhaps not surprisingly, we built a model showing how GSIM might better achieve its objectives.

47. The Generic Model Value Pyramid (see Figure 2) has four layers: Language, Community, Work and Results. These four layers contain the six GSIM objectives: Communicate; Educate; Collaborate; Validate; Innovate and Automate. Starting from the bottom of the pyramid, each layer is a prerequisite for the layer above.

48. Before you get Results by Automating your statistical business, you Work to Validate that you are doing so based on existing best practice or, when this is not sufficient, by Innovating a "new" best practice. In order to do effective Work, you need to build a Community by Educating people across functions (and possibly organizations) so they can Collaborate on validating, innovating and, ultimately, automating. Finally, it is impossible to build a strong Community without Communicating using a shared Language.

49. Of course, you have to build the pyramid from the bottom up:

1. Use Language to Communicate
2. Build Community by Educating and Collaborating
3. Do the Work to Validate and Innovate
4. Get Results by Automating

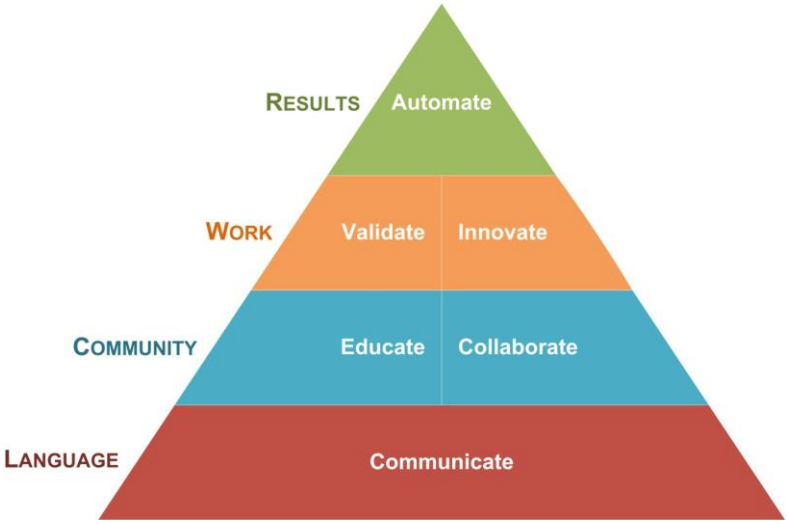

Figure 2 – Generic Model Value Pyramid.

Use Language to Communicate

50. Lexicographers do not build languages. Instead, lexicographers record language used by speakers and writers. By doing this, dictionaries stay relevant, while providing a stable and common reference point.

51. GSIM provides a language for users of official statistics information objects. Those developing GSIM are lexicographers of its dictionary. We need to reflect the language of official statistics, and draw out consensus to form a common reference point for statisticians and related professionals.

52. We should not define language for every possible object, but rather define existing terminology as common reference points for GSIM users. The GSIM objectives anticipate a broad constituency, including not only staff of statistical offices, but also users of official statistics.

53. We think GSIM v1.0 goes too deep, with objects being too technical for the GSIM lexicon at this stage of its maturity. To meet the objective to communicate, we feel the model needs fewer objects overall, and these objects should be at a higher level.

Build Community by Educating and Collaborating

54. Successful communities grow by building a body of people who understand its principles, talk its language and seek to belong. Education drives growth, as evidenced by the success of a vast range of missionaries over the course of human history.

55. Successful education connects to people's everyday life. It is facilitated by creating a "simple, easy to understand view of complex information, with clear definitions."[4]

56. Creating an easily understood view also helps people collaborate. Working with others using a shared language and a common understanding helps participants feel they belong to a group much bigger than their work unit, their project team, or even their organization.

57. With GSIM, potential users must "buy-in" at the most detailed level of the model. There is no rigorous, agreed and coherent picture at any level above the specification, which, at nearly 300 pages, is not a usable tool for educating or collaborating.

58. We believe GSIM needs a model of about 50 high level objects as its main organizing scheme. This model would stand as an analogue to the GSBPM model of sub processes. The two could operate in tandem as tools to educate and collaborate.

[4] The Generic Statistical Information Model (GSIM) brochure, page 2: http://www1.unece.org/stat/platform/display/metis/Brochures.

Do the Work to Validate and Innovate

59. While communicating, educating and collaborating create the capabilities and capacity that form the basis of GSIM work, validation and innovation are the engines generating the model's value.

60. Validation enables teams to quickly test how well existing processes work, in order to identify processes and objects to improve. This exercise leads teams to innovate locally, by adopting best practice where comparison against the generic model showed a gap. We can expect broader moves to innovate once GSIM leads to more widespread efforts to collaborate. An effective GSIM will help good ideas transmit much faster due to the common language and understanding surrounding the model.

61. Our work to validate the IMF Statistics Department's IMD product was conducted by a group of staff familiar with the development of GSIM, not the IMD project team. We intended to involve them in the work, but it quickly became evident staff found the GSIM model too complicated and too detailed to gain enough understanding within a reasonable timeframe. Instead, we translated the more easily understood GSBPM model into questions seeking answers about the objects used in the IMD process.

62. This approach is not scalable. While we hope to continue to use GSIM, it will remain the domain of metadata and process experts, rather than being adopted more widely.

63. We can address this problem in the same way the educate and collaborate challenges can be overcome, by creating a higher level object model, to stand as a complete and coherent story of the main objects, and their relationships, used in the statistics process.

Get Results by Automating

64. Ultimately, GSIM (working with GSBPM) will succeed when it helps us automate manual or partially automated processes, leading to reduced cost and improved quality, helping statistical offices expand the range of sources and volume of data they can process.

65. The fine details contained in the GSIM model start to provide significant value at the automate part of the Value Pyramid. The question to consider is if these details should form part of GSIM itself, or if they should form the basis of an ancillary implementation standard.

66. A significant amount of study and analysis is needed before arriving at a decision. Whatever the outcome, the overall GSIM scheme must reflect that the objects needed to help automate processes are a level deeper than those needed for meeting the other GSIM objectives.

D. GSIM Level 2

67. Whether we have a high-level GSIM with an ancillary detailed implementation standard or a multi-level GSIM with different levels of detail, we find a very strong need to have a complete and coherent object model at a higher level than GSIM v1.0 possesses. Such a model would form GSIM Level 2, Level 1 being the existing four Groups and Level 0 being the statistical information objects class. (Note that we advise reviewing Level 1 as part of any process to create a Level 2 of the model.)

68. GSIM Level 2, analogous to the GSBPM sub-processes level (also Level 2), would contain around 50 objects and model their most common relationships. It would provide object descriptions and indicate sub-types or other more detailed implementations of the objects. Relationships to the more detailed objects could be provided in the same free text way sub-processes are described in Level 3 of GSBPM. Rigorous links to lower levels or ancillary standards could follow in later GSIM versions.

69. We suspect most of the information objects required for a Level 2 model already exist within the GSIM model, although some new ones may need to be created to combine or summarize multiple existing objects. However, we need to do considerable work to reach a coherent and easily communicated model between objects at this level, and to create accessible descriptions of each of these objects. We encourage the GSIM work program (in which the IMF is an active participant) to incorporate this work as a matter of urgency.