GSIM and more

InKyung Choi
United Nations Economic Commission for Europe (UNECE)

@SESRIC Workshop July 10th 2023
Agenda

1. Introduction to GSIM
2. Linking GSBPM and GSIM
3. SDMX-DDI-GSBPM
Part 1: Introduction to Generic Statistical Information Model (GSIM) *

* This presentation materials are based on “Introduction to GSIM” by Jenny Linnerud (Norway) from GSIM E-training
Modernisation

• Modernisation of statistics can be supported by re-use and sharing of methods, components, processes, metadata and data

• Generic Statistical Information Model (GSIM) is the first internationally endorsed reference framework for statistical information

• GSIM is a conceptual framework playing an important role in modernising, streamlining and aligning the standards and production associated with official statistics, at both national and international levels.
What is GSIM?

• It is a reference framework of internationally agreed definitions, attributes and relationships that describe the pieces of information used in the production of official statistics (called “information class”).

  - Subject Field
    - Definition: field of knowledge under which a set of Concepts and their Designations is used

  - Statistical Programme
    - Definition: set of activities to produce statistics on a given Universe within the context of Subject Fields
    - Attributes: Budget, Source of Funding, etc.

  - Statistical Programme Cycle
    - Definition: iteration of a Statistical Programme for a given Population
    - Attributes: Reference period, etc.

• This framework enables generic descriptions of the definition, management and use of data and metadata throughout the statistical production process.
Development of GSIM

- Developed by statistical organisations and for statistical organisations
- GSIM version 1.0 (2012)
- GSIM version 1.1 (2013)
- GSIM version 1.2 (2019)
- GSIM version 2.0 (2023)

17 statistical organisations
GSIM Groups

- User needs
- The structure of a statistical program, process steps, methods, rules
- Execution of the process

- How is data structured, what format is it in, where is it stored
- Reference metadata - documention

Reusable attributes for information classes, identity, versioning etc

The information that comes in and out of the organisation, exchange instrument, provision agreements.

Statistical content - definitions of concepts, variables, populations, unit types
The meaning of codes; statistical classifications, code lists
GSIM Groups

- The **Concepts** group is used to define the meaning of data, providing an understanding of what the data are measuring.
- The **Structures** group is used to describe and define the terms used in relation to information and its structure.
- The **Business** group is used to capture the designs and plans of statistical programs, and the processes undertaken to deliver those programs. This includes the identification of a *Statistical Need*, the *Business Processes* that comprise the *Statistical Programme* and the Assessment of them.
- The **Exchange** group is used to catalogue the information that comes in and out of a statistical organization via *Exchange Instruments*. It includes information classes that describe the collection and dissemination of information.
Some GSIM information classes

127 information classes exist in GSIM (version 2.0)
GSIM information classes that tell a story
What is the relationship between GSIM & GSBPM?

- GSIM and GSBPM are complementary models for the production and management of statistical information. Greater value can be obtained when both models are applied together.
- GSBPM models the statistical production process and identifies the activities undertaken by producers of official statistics that result in information outputs.
- GSIM helps describe GSBPM sub-processes by defining the information classes that are used by them, that flow between them, and are created in them in order to produce official statistics.
GSIM experiences

Statistics Canada from GSIM e-training (2019)

Istat (Italy) from GSIM e-training (2019)

Hungary from ModernStats World Workshop (2022)

Sweden from ModernStats Community of Practice (2023)

Finland from ModernStats World Workshop (2019)

France from ModernStats World Workshop (2022)

GSIM objects pinpointed

- GSIM is rich enough to provide useful objects.
- Statics dimension: to characterise the main steps or objects of the process
  - Business services, Process Steps, Business Functions. All of them tagged with attributes and administrative details.
- Dynamic dimension: to describe the sequence of steps
  - Objects around the process design
    - input (core, process support, parameters)
    - output (core, process execution log)
  - Method (specification of the technique used in the process)
  - Roles (mathematical or logical expression)
Benefits of GSIM experienced

• Provides a common language to improve communication between statistical organisations at national and international levels
• GSIM in combination with GSBPM creates an environment prepared for reuse and sharing of methods, components and processes
• Enables statistical organisations to rethink how their business could be more efficiently organised by defining information classes common to all statistical production, regardless of subject matter
• Provides a common language to improve communication between the different roles in statistical production (business and information technology experts)
Part 2: Linking GSBPM-GSIM*

* This presentation materials are based on “Linking GSBPM and GSIM” by Flavio Rizzolo (Canada) from ModernStats World Workshop 2022
Linking GSBPM and GSIM Task Team (2019-21)

- 20 members from 11 participating institutions (Australia, Canada, Egypt, Hungary, Italy, Mexico, Poland, South Korea, Sweden, ILO, DDI)
- Derived a robust set of GSIM classes that could be used as inputs and outputs based on use cases commonly taking place in many statistical organisations.
- Made it easier to design systems to track information flow through statistical business processes
- Contributed to building a “de facto” integrated view of the ModernStats models (now expanded by the Core Ontology for Official Statistics work)
- Made it easier to create implementations, in particular CSPA services (now continued by the SDMX-DDI-GSBPM mapping work)
Linking GSBPM and GSIM - Context

GSIM information class

GSBPM sub-process

GSIM information class

> 120 classes

> 40 sub-processes

> 120 classes
Linking GSBPM and GSIM - Specification

GSIM **Business** group is used to capture the designs and plans of statistical programmes, and the processes undertaken to deliver those programmes.
Linking GSBPM and GSIM - Specification

Core Input
- Data Set

Process Support Input
- (auxiliary) Data Set

Parameter Input

Core Output
- (updated) Data Set

GSBPM sub-process specification
- Process Step
- Process Method
- Control Design
- Process Design

Process Metric
- Process Execution Log
## Specification example: GSBPM sub-process 5.3 Review and Validate

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Core Input type</strong></td>
<td><strong>Process Method</strong></td>
<td><strong>Core Output type</strong></td>
</tr>
<tr>
<td>Data Sets (unit Data Sets) to be reviewed and validated</td>
<td>Review Data Sets and Process Methods</td>
<td>Data Sets (unit Data Sets): updated Data Set</td>
</tr>
<tr>
<td>Data Structures associated with Data Sets to understand Data Sets</td>
<td>Apply Process Methods and Rules to review Data Sets</td>
<td>Data Structure associated with Data Set</td>
</tr>
<tr>
<td>Represented Variables to be reviewed and validated</td>
<td>Apply Process Methods and Rules to validate Data Sets</td>
<td>Referential Metadata Set: descriptions of the Process Methods used, quality information summarising Process Metrics or any other relevant information to be passed along with Data Sets</td>
</tr>
<tr>
<td>Process Methods that specifies methodology for review and validation (e.g., calculating plausibility or validity) which can be represented as Rules, as designed in Phase 2</td>
<td>Calculate quality measures specified by Process Methods</td>
<td><strong>Process Metric type</strong></td>
</tr>
<tr>
<td><strong>Parameter Input type</strong></td>
<td><strong>Process Support Input type</strong></td>
<td><strong>Process Execution Log type</strong></td>
</tr>
<tr>
<td>Parameter values to be used for review and validation methodologies as specified in Process Method such as:</td>
<td>Auxiliary Data Sets or any Information Resource to be used for review and validation, e.g., historic comparison, macro-level comparison</td>
<td>Execution log such as</td>
</tr>
<tr>
<td>• Limit value for edit Rule (interval for valid values)</td>
<td>Technical / methodological handbooks, policies or guidelines to be followed regarding data validation as well as quality management</td>
<td>• Time that Process Step started</td>
</tr>
<tr>
<td>• Threshold for checking outlier</td>
<td></td>
<td>• Time that Process Step ended</td>
</tr>
</tbody>
</table>

### Process Design
- **Process Method**
  - Review Data Sets and Process Methods
  - Apply Process Methods and Rules to review Data Sets
  - Apply Process Methods and Rules to validate Data Sets
  - Calculate quality measures specified by Process Methods
  - Update Data Sets and associated element in Data Structure with results from review and validation

### Process Output Specification
- **Core Output type**
  - Data Sets (unit Data Sets): updated Data Set
  - Data Structure associated with Data Set
  - Referential Metadata Set: descriptions of the Process Methods used, quality information summarising Process Metrics or any other relevant information to be passed along with Data Sets

- **Process Metric type**
  - Quality measures related to review and validation such as:
    - Number of validations conducted
    - Number of outliers detected
  - Quality measures of Process Step such as:
    - Time spent to complete the Process Step (derived from Process Execution Log)
    - Cost spent to complete the Process Step

- **Process Execution Log type**
  - Execution log such as
    - Time that Process Step started
    - Time that Process Step ended
    - Any message or event log generated from software used for review and validation
GSIM Exchange Group in Action

Scenario: NGO and a sewage treatment company recently established Memorandum of Understanding (MoU) that agrees on the principle of data sharing. The company commits to make their data available via API for NGO to use for statistical purpose.

This diagram aims to demonstrate how GSIM classes in Exchange Group can be used to represent Information in GSVP in sub-processes relevant to exchange (while the focus of this diagram is on Exchange Group, some of non-Exchange Group classes were also added to make the story flow more smooth.)
About this diagram: this diagram aims to demonstrate how GSIM classes in Business Group can be used to represent information in GSIPM phrases / sub-processes. The goal is NOT to show all and every GSIM classes needed but to showcase how GSIM classes can be used in the context (while the focus is on Business Group). Some of the non-Business Group classes were also added to make the story flow more smoothly.

Scenario (hypothetical) used in the diagram: NSO and a sewage treatment company recently established Memorandum of Understanding (MoU) that agrees on the principle of data sharing. The company agreed to make their data available via API for NSO to use for statistical purposes.
GSIM classes for all GSBPM sub-processes

Available on the Linking GSBPM-GSIM “information flow” wiki page
Part 3: GSBPM-SDMX-DDI

* This presentation materials are based on “Supporting Standards Group” by Zoltán Vereczkei (Hungary), Chair of the Supporting Standards Group, from Modernisation Workshop 2022
GSBPM-SDMX-DDI Task Team Context

Benefits
- International / global standards
- Free / open
- Helps reuse tools and concepts
- Increases interoperability
- Improves quality

Difficulties
- Which standard to use for which use case?
- Technical, people get lost
- Which artefact to use for which stage of production process?
GSBPM-SDMX-DDI Task Team (2022-)

• 15 members from 11 participating institutions (Canada, Hungary, Italy (NSO), Italy (Bank), Mexico, Kingdom of the Netherlands, USA, BIS, DDI, ILO, OECD)

Statistical Data and Metadata eXchange
Data Documentation Initiative

Provides statistical context
Provides inputs / outputs that can be mapped to SDMX / DDI artefacts
GSBPM-SDMX-DDI Task Team Objectives

Source: Steven Vale: Exploring the relationship between DDI, SDMX and the Generic Statistical Business Process Model
GSBPM-SDMX-DDI Task Team Objectives

Objectives
- Provide short description why/how SDMX/DDI helps as an entry point
- List relevant SDMX/DDI artefacts under each sub-process
- Map relevant SDMX/DDI artefacts under each sub-process to GSIM class

Results and (some preliminary) findings
- By using GSBPM as common linkage points, identifies where SDMX and DDI work together, their strengths
- With mapping between SDMX and DDI made easier, more chances for interoperability between the two standards

Example for GSBPM sub-process 1.4
Resource

- GSIM Wiki
- GSIM Resource Repository
- GSIM E-training (2019)
- Linking GSBPM-GSIM
- GSIM implementation roadmap and tools
Thank you!

Feel free to contact me for more information (choii at un dot org)