New Developments from the HLG-MOS Supporting Standards Group - Linking GSBPM and GSIM -

Flavio Rizzolo (Statistics Canada) - June 28th, 2022
Standards landscape

Sub-processes (GSBPM)

Information (GSIM)

Structures Concepts

Capabilities (CSDA)

Activities (GAMSO)

PROV-O

...and many more...

DCAT

SKOS

services (CSPA)
Linking GSBPM and GSIM - Task Team

20 members over 3 years from 11 participating institutions

<table>
<thead>
<tr>
<th>Country</th>
<th>Member</th>
</tr>
</thead>
<tbody>
<tr>
<td>Italy</td>
<td>Marina Signore - Chair Giorgia Simeoni</td>
</tr>
<tr>
<td></td>
<td>Mauro Scanu</td>
</tr>
<tr>
<td></td>
<td>Claudia Brunini Carlo Vaccari</td>
</tr>
<tr>
<td>Canada</td>
<td>Flavio Rizzolo - Chair Francine Kalonji</td>
</tr>
<tr>
<td>Mexico</td>
<td>Juan Munoz Ildeliza Ramos Jesús Togno</td>
</tr>
<tr>
<td>Sweden</td>
<td>Patrik Wahlgren</td>
</tr>
<tr>
<td>Hungary</td>
<td>Zoltán Vereczkei Csaba Ábry Eva Hajosn Ender</td>
</tr>
<tr>
<td>Poland</td>
<td>Anna Dlugosz</td>
</tr>
<tr>
<td>South Korea</td>
<td>Yulla Choi</td>
</tr>
<tr>
<td>Australia</td>
<td>Annette Tyler</td>
</tr>
<tr>
<td>ILO</td>
<td>Edgardo Greising</td>
</tr>
<tr>
<td>DDI</td>
<td>Jay Greenfield</td>
</tr>
<tr>
<td>Egypt</td>
<td>Ayman Hathoot</td>
</tr>
</tbody>
</table>
Linking GSBPM and GSIM - Goals

• Derive 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.
• Elicit non-obvious relationships between the two models, improving usability and supporting a wider adoption;
• Make it easier to design systems to track information flow through statistical business processes;
• Contribute to building a “de facto” integrated view of the ModernStats models (now expanded by the Core Ontology for Official Statistics work);
• Make it easier to create implementations, in particular CSPA services (now continued by the just-started SDMX-DDI-GSBPM mapping work)
Linking GSBPM and GSIM - Context

GSIM class \(\rightarrow\) GSBPM sub-process \(\rightarrow\) GSIM class

- \(> 100\) classes
- \(> 40\) sub-processes
- \(> 100\) classes
Linking GSBPM and GSIM - Specification

Core Input Type:
- GSIM class

Parameter Input Type:
- GSIM class

Process Support Input Type:
- GSIM class

Core Output Type:
- GSIM class

Process Metric Type:
- Process Execution Log Type

GSBPM sub-process specification:
- GSIM Process Step
- GSIM Process Method
- GSIM Control Design
- GSIM Process Design
Linking GSBPM and GSIM - Instance

Core Input
- GSIM instance

Parameter Input

Process Support Input
- GSIM instance

GSBPM sub-process instance
- GSIM Process Step Instance
- GSIM Process Control

Core Output
- GSIM instance
- Process Metric
- Process Execution Log
Specification template example: 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>Update Data Sets and associated element in Data Structure with results from review and validation</td>
<td>Quality measures related to review and validation such as:</td>
</tr>
<tr>
<td>Parameter values to be used for review and validation methodologies as specified in Process Method such as:</td>
<td></td>
<td>• Number of validations conducted</td>
</tr>
<tr>
<td>- Limit value for edit Rule (interval for valid values)</td>
<td></td>
<td>• Number of outliers detected</td>
</tr>
<tr>
<td>- Threshold for checking outlier</td>
<td></td>
<td>Quality measures of Process Step such as:</td>
</tr>
<tr>
<td><strong>Process Support Input type</strong></td>
<td></td>
<td>• Time spent to complete the Process Step (derived from Process Execution Log)</td>
</tr>
<tr>
<td>Auxiliary Data Sets or any Information Resource to be used for review and validation, e.g., historic comparison, macro-level comparison</td>
<td></td>
<td>• Cost spent to complete the Process Step</td>
</tr>
<tr>
<td>Technical / methodological handbooks, policies or guidelines to be followed regarding data validation as well as quality management</td>
<td></td>
<td><strong>Process Execution Log type</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Execution log such as:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Time that Process Step started</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Time that Process Step ended</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Any message or event log generated from software used for review and validation</td>
</tr>
</tbody>
</table>
Linking GSBPM and GSIM - Information Flows

GSBPM 4.4 
Finalize Collection

GSBPM 2.5 
Design Processing 
and Analysis

GSBPM 5.4 
Edit & Impute

GSBPM 5.3 
Review & Validate

Reference 
Documents 
(for E&I)

Represented 
Variable (collected 
to be E&I'ed)

Data Structure 
(collected to 
be E&I'ed)

Data Set 
(collected to 
be E&I'ed)

Process 
Method 
(for E&I)

Reference 
Documents 
(for R&V)

Represented 
Variable (E&I’ed 
to be R&V’ed)

Data Structure 
(E&I’ed to be 
R&V’ed)

Data Set 
(E&I’ed to 
be R&V’ed)

Process 
Method 
(for R&V)

Represented 
Variable (R&V’ed)

Data Structure 
(R&V’ed)

Data Set 
(R&V’ed)

Reference 
Metadata Set 
(E&I summaries)

Referential 
Metadata Set 
(R&V summaries)

Referential 
Metadata Set 
(E&I summaries)
Outcomes and lessons learned

• The two templates – specification and instance (running examples) – provide a useful way of eliciting use cases from participating NSOs

• The new information flow diagrams (https://bit.ly/3OH2pwb) show which GSIM classes are most used across the statistical business process… and the gaps that exist


• Useful feedback to the GSIM revision and Core Ontology for Official Statistics task teams

• Mapping exercise proved to be more challenging for early GSBPM phases
  • Processes around requirements and design are less clear
  • Terminological issues between the two models became more problematic
  • Subject matter expertise was harder to find

• A mapping to implementation standards, e.g. DDI and SDMX, will help making the models more precise and interoperable -> new task team on SDMX-DDI-GSBPM mapping
Related Links

• Generic Statistical Information Model (GSIM): https://bit.ly/3I8SA8x
• New GSIM Revision: https://bit.ly/3OEDnxO
• GSBPM-GSIM information flows and report: https://bit.ly/3Ov1JdW
• Core Ontology for Official Statistics (COOS): https://github.com/linked-statistics/COOS
ModernStats World Workshop 2022

27-29 June @Belgrade, Serbia

Twitter: @unecestat #ModernStats
Wiki page for timetable and files: https://statswiki.unece.org/x/sYAqF