Towards Practical Interoperability: Mapping SDMX and DDI for Data Integration

Edgardo Griesing (ILO), **Matjaz Jug** (Statistics Netherlands)
and Flavio Rizzolo (Statistics Canada)

(on Behalf of the UNECE Supporting Standards Group)

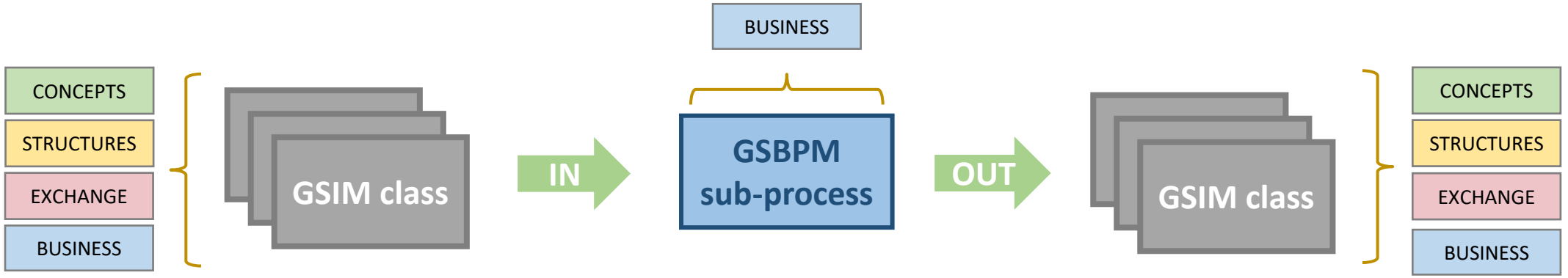
Data integration use cases

- Data Integration: ability to combine, link, relate and different in various ways for data production, analysis, understanding complex phenomena, etc.
- Large number of use cases in multiple domains (and across domains)
  - UN Statistical Division SDGs SDMX data -> Google Data Commons via DDI-CDI
  - NADA data catalogs (IHSN) with 1000's of studies in DDI-Codebook -> SDMX via DDI-CDI
  - A growing number of WorldFAIR (and now WorldFAIR+) and CDIF projects

Mapping approach

- DDI-CDI comes in handy as a *Rosetta stone* for mappings and integration
  - Focuses on multidisciplinary (and interoperable) data sharing and integration
  - Provides multiple syntax representations for machine-actionability
  - Complements (and integrates with) other DDI products
- Prerequisites for effective data integration
  - Include *reference* and *structural metadata* to describe data well enough to make it integration-ready (including structure, semantics and codesets)
    - Define mappings between concepts, variables and codesets as necessary (expressed in SKOS/SSSOM)
    - Harmonize concepts/variables/formats (whenever possible) or document the caveats (if not 100% possible)
  - Preserve *process metadata* for provenance (description of operations performed on the data)

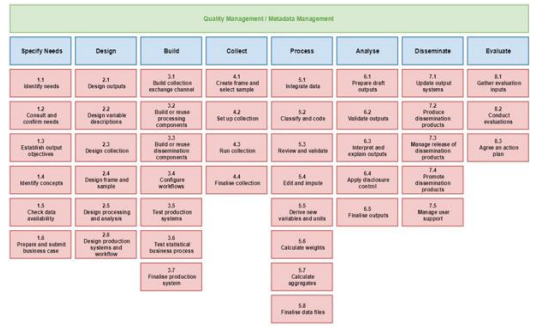
Building interoperable GSBPM pipelines



- Surface non-obvious relationships between the two models, improving usability and supporting a wider adoption
- Select a robust set of GSIM classes that could be used as inputs and outputs of GSBPM sub-processes based on use cases commonly taking place in many statistical organisations
- Help in the design of systems that track information flows through statistical business processes **by reusing structural metadata (concepts, structures..)**
- **Make it easier to build interoperable pipelines with popular implementation standards, e.g. SDMX, DDI, etc**.

Interoperability using SDMX and DDI with GSBPM

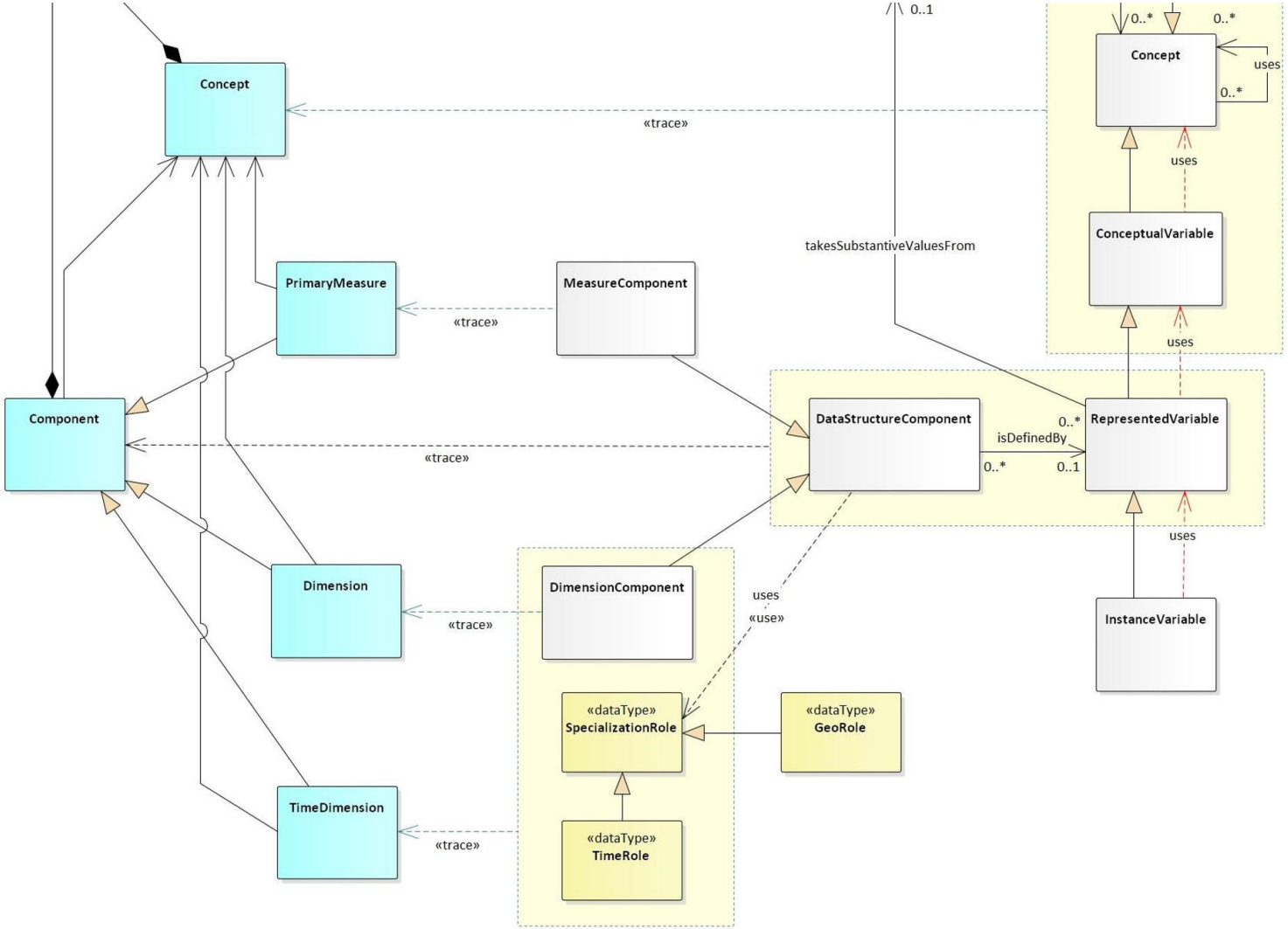
- **Benefits**
  - International standard suites with large adoption base
    - Include DDI Codebook, DDI Lifecycle, DDI CDI, SDMX 2.1 and 3.x, XKOS, VTL and others.
  - Open and collaborative development
  - Growing ecosystem of tools and libraries readily available (key for efficient development)
  - Support for structural and semantic harmonization and alignment
- **Issues**
  - Developed independently by difference communities with little intentional alignment
  - Impedance mismatch: varying emphasis on semantics, structures and granularity
  - Different strengths (and weaknesses)



Results

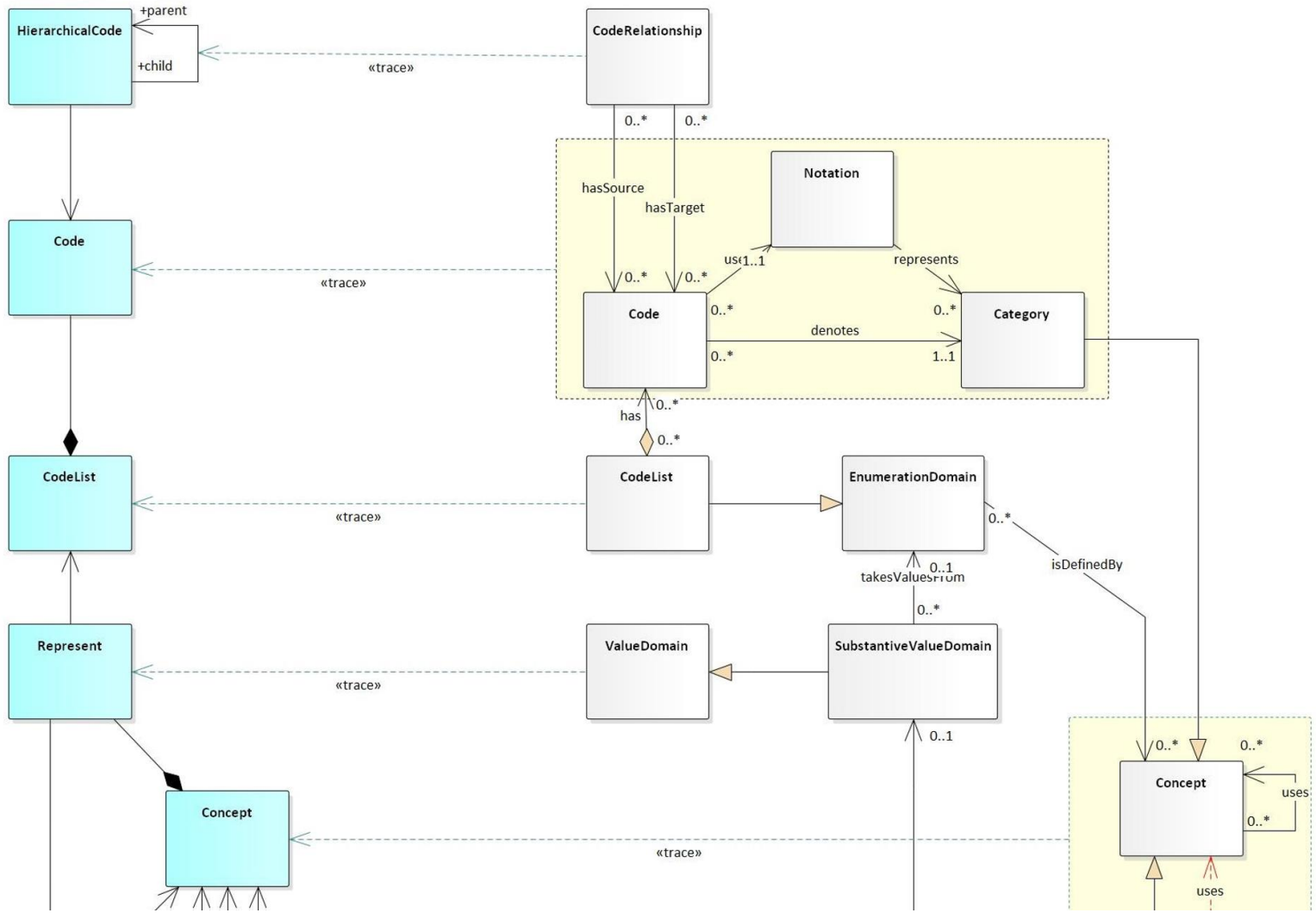
- Identified relevant SDMX and DDI artefacts for each GSBPM sub-process (**finished**)
- Developed guidelines on how to make both standards interoperate with each other and with VTL to enable data production pipelines for statistical production
- Developed foundation for further work to include SDTL, XKOS, and other open implementation standards

Contextual mapping



- Mappings are many-to-one and many-to-many
- The usage of a class depends on the related classes (context)
- The same class can be mapped differently in different contexts (and for different use cases)

Contextual mapping (cont.)



Expected outputs

- Guidelines based on use cases
  - No prescriptive framework given the potential number of scenarios: we cannot know how exactly a user will work with the data for integration purposes
- Data description elements (from DDI and SDMX) required to make the data integration-ready
- Machine-actionable mappings based on standards (SKOS/SSSOM, others)
- Clarification of potentially confusing terminology
  - DDI and SDMX tend to use similar terms for notions that are not exactly the same, e.g. Concepts, Categories, Codes, Components, etc. whereas some other notions are explicitly defined in one standards but not in the other, e.g. Variable.
- Recommendations on how to define the integration process based on mappings and data structures
- Identification of potential interoperability with other standards (DCAT, schema.org, etc.)

ModernStats models and architectures provide the conceptual framework to better understand where (and how) to use DDI and SDMX