Why SDMX matters? A community journey towards SDMX as an AI-enabler and a Data Mesh enabler

SDMX Global Conference, Empowering Data Communities
Manama, October 2023

Keywords
#OpenSource #OpenKnowledge #Community #SDMX #MetadataDriven #AI #DataMesh
1. Creating value with #SDMX – A #Community journey

2. The #SDMX #MetadataDriven approach

3. #SDMX as #AI and #DataMesh enabler?
CREATING VALUE WITH #SDMX
A #COMMUNITY JOURNEY
Who we are

SIS-CC (“the Community”) is a reference open-source community for official statistics, focusing on product excellence and delivering concrete solutions to common problems through co-investment and co-innovation.
#SDMX #OpenSource – The .Stat Suite

- **.Stat Core**
  - ...to store data accessible through SDMX APIs

- **.Stat Data Explorer**
  - ...to explore data from any SDMX endpoint(s)

- **.Stat Data Lifecycle Manager**
  - ...to manage data and metadata through the lifecycle
#SDMX enable a cooperative “Business Model”
Examples of implementations
#SDMX #OpenKnowledge – The .Stat Academy

https://academy.siscc.org

Starting point: Build capacity in SDMX data modelling
THE #SDMX METADATA-DRIVEN APPROACH
The #SDMX Backbone

SOURCE & COLLECT (META)DATA

PROCESS & ANALYSE (META)DATA

DISSEMINATE DATA & ENGAGE

MANAGE (META)DATA

MANAGE DATA PROJECT & ALGORITHMS
The #SDMX Backbone

SOURCE & COLLECT (META)DATA

MANAGE (META)DATA

PROCESS & ANALYSE (META)DATA

DISSEMINATE DATA & ENGAGE

MANAGE DATA PROJECT & ALGORITHMS
The #SDMX Backbone

The “Data Integration” use case at the National Bank of Belgium

Source: SALSA Project – National Bank of Belgium
The #SDMX Backbone – Wrap-up

- SDMX data storage becomes a commodity
- SDMX APIs enable universal data interoperability
- SDMX semantics enables “metadata-driven” data user / manager experience
Step 1 Home page
Lexical search or Topics

Step 2 Search results
Filtering on dynamic facets

Step 3 Visualisation page
Selecting dimensions / variables

#SDMX #MetadataDriven Data User Experience

category scheme(s)
dimensions: concepts + codelists
dataflow + presentation annotations
#SDMX #MetadataDriven Data **User** Experience

Data Manager → Information Model → Data User

User-centered design  
Data harmonisation
Establish a systematic User Research
The Data Platform is fragmented

<table>
<thead>
<tr>
<th>Domain-Specific Data System</th>
<th>Domain-Specific Data System</th>
<th>Domain-Specific Data System</th>
</tr>
</thead>
<tbody>
<tr>
<td>Domain specific (meta)data model</td>
<td>Domain specific (meta)data model</td>
<td>Domain specific (meta)data model</td>
</tr>
<tr>
<td>Manual editing, Validations reports &amp; Quality checks</td>
<td>Manual editing, Validations reports &amp; Quality checks</td>
<td>Manual editing, Validations reports &amp; Quality checks</td>
</tr>
<tr>
<td>Domain specific data pipelines</td>
<td>Domain specific data pipelines</td>
<td>Domain specific data pipelines</td>
</tr>
<tr>
<td>Domain specific Statistical functions</td>
<td>Domain specific Statistical functions</td>
<td>Domain specific Statistical functions</td>
</tr>
</tbody>
</table>

**Data models embedded deep in DBs, difficult to extract, map or reuse**

**Limited capacity for automation, predominantly manual interventions**

**Duplications and inconsistencies in the statistical methods, plus limited reproducibility**

**Domain-specific systems prone to obsolescence and lack reusability**

**Vertical skills silos difficult to maintain and harms mobility**
# SDMX #MetadataDriven Data Manager Experience

## Metadata-driven pipelines
Manipulating SDMX data
- Extraction, Harmonisation, Chained calculations, Derived indicators

## Metadata-driven statistical functions
Manipulating SDMX data
- Interpolation, Chain linking, Aggregation functions, Seasonal adjustments, Smoothing, Rebasement, Backcasting, Forecasting, Growth rates

### (Meta)Data Management
- Structural Metadata
  - Categories, DSDs/MSDs, Dataflows
- Referential Metadata
  - Data quality, Methodology
- Process Metadata (?)
  - Extraction, Transformation Metadata

### Data Engineering
- Validations reports & Quality checks

### Data Transformation
- Data Analysis
  - Data Product
    - Discoverable, Understandable, Interoperable, Globally addressable

### Data Analysis
- Power BI

---

(?) feature that exists partially / peacemeal in SDMX and that ought to be specified
The #SDMX #MetaDataDriven approach combines (centrally managed) data infrastructure with the possibility of (decentralized) data operations, creating multiple benefits:

- Decoupled roles and specializations
- Data mgt by non coding-experts
- Facilitation of staff mobility
- Possibility of data tasks mutualization
- Facilitation of platform migration
- Domain specific coding language
- Domain specific interfaces
- Capitalize on shared methods / reuse
- Capitalize on common infrastructure
Structure your Data governance
SDMX provides the tools (conceptual and technical) to operate your governance and progressively harmonise information models.
#SDMX #MetadataDriven approach – Wrap-up

- Systematic user research – enabled by SDMX
- Structured data governance – enabled by SDMX
- SDMX data storage becomes a commodity
- SDMX APIs enable universal data interoperability
- SDMX semantics enables “metadata-drivenness”

Drive
Data User experience
Data Manager experience
With SDMX metadata
#SDMX #AI
SDMX AS AN AI ENABLER?
Opportunities #SDMX #AI

Nowcasting
Federated Learning

Outlier detection
Clustering
Data Curation

Natural Language
Conversational
Data Access

Pair programming
Pair modelling
The cognitive gap
Develop a natural language capable Data Explorer (I)

How many booster doses of covid vaccines administered in New Caledonia in 2023?

Step 2.5 or 3.5
Natural Language query transformed into search queries
+ Pre-selection of dimensions on target dataflow(s)
Develop a natural language capable Data Explorer (I)

**Proof of Concept ongoing, mixing Generative AI & SolR techniques.**

- Natural language query
- List of all topics and dimensions returned (SolR index wrapping sdmx structures)
- List of additional reformulations (openAI *generation* of synonyms and related terms)
- List of most relevant topics and dimensions (openAI *extraction* of topics and dimensions)
- List of dataflows SolR query (generated reformulations, extracted topics/dimensions)
- List of dimension values (openAI *extraction* of dimension values)
- JSON with selected dataflows and dimension values

Source: courtesy by A. Benedetti, SEASE.
The universal StatsBot to access any SDMX source (II)

The StatsBot is within reach... we could develop it as an open and reusable asset.

Source: courtesy by K. Castillo Perez, Expertime.
Develop a Natural language capable Data Explorer (I)

Develop the universal StatsBot to access any SDMX source (II)

Develop a “prompt engineering in an SDMX context” practice

Develop the specialized, “SDMX-literate” open source LLM
#SDMX #DATAMESH
SDMX AS A DATA MESH ENABLER?
What is a #DataMesh?

Address the problem with the centralized organization behind the DWH/Data Lake approach to analytical data: centralized expertise often leads to bottleneck and failure.

The #DataMesh is the socio-technical paradigm to switch from a purely centralized model to a model that enables domain-oriented decentralization of analytical data operations.

Source: Data Mesh Architecture (datamesh-architecture.com)
“Data mesh attempts to strike a balance between team autonomy and interteam interoperability and collaboration. It gives domain teams autonomy to take control of their local decision making, for example, choosing the best data model for their data products, while it uses computational governance policies to impose a consistent experience across all data products, for example, standardising the data modelling language that all domains utilize. Data mesh gives domain teams autonomy to build and maintain their data products, while it places a domain-agnostic data platform in place for teams to do so in a consistent and cost-effective way.”

Zhamak Dehghani, Data Mesh
#SDMX as a #DataMesh enabler?

Source: Data Mesh Architecture (datamesh-architecture.com)
SDMX governance as cornerstone of data governance, enabling a progressive harmonisation of data semantics across domains and the data lifecycle.

Each business domain to build their data process and products in relative autonomy, drawing on common information models and technical building blocks.

SDMX backbone brings commoditised SDMX storage, SDMX API-centricity, and SDMX semantics as part of the “Data infrastructure as a Platform”.

Source: Data Mesh, O’Reilly, Z. Dehghani
#SDMX #AI Co-investment and co-innovation opportunities

- Develop implementation guidelines for the #SDMX #DataMesh & Experiment
- Develop the architecture vision & business case for #DataMesh in Official Statistics
- Adapt and augment #SDMX (and #VTL?) with the #DataMesh paradigm (data product, data contract, DDD, computational governance...)
- Develop implementation guidelines for the #SDMX #DataMesh & Experiment
Reference resources

SIS-CCommunity

• SIS-CC Community
• SIS-CC 2020-25 Strategy
• Why SDMX Matters

SDMX.IO Ecosystem

• The SDMX.IO ecosystem

SDMX Community

• SDMX Community
• SDMX 2021-25 Strategy
• The business case for SDMX
• SDMX 3.0 packages under review
• SDMX Guideline: Modelling statistical domains and exchange frameworks in SDMX
• SDMX Guidelines: A Reference Framework for SDMX Structural Metadata Governance

.Stat Suite, open source platform

• .Stat Suite – High level product overview and flight planner
• .Stat Suite – Kanban Board
• .Stat Suite – Changelog
• .Stat Suite – Extensive documentation
• .Stat Suite – Code repository

.Stat Academy, open knowledge platform

• .Stat Academy – Learning Paths for Data Producer; Data Tooler
• .Stat Academy – Courses
• .Stat Academy – Webinars
• .Stat Academy – Become an instructor
1. Creating value with #SDMX – A #Community journey

2. The #SDMX #MetadataDriven approach

3. #SDMX as #AI and #DataMesh enabler?

co-investment

Co-innovation