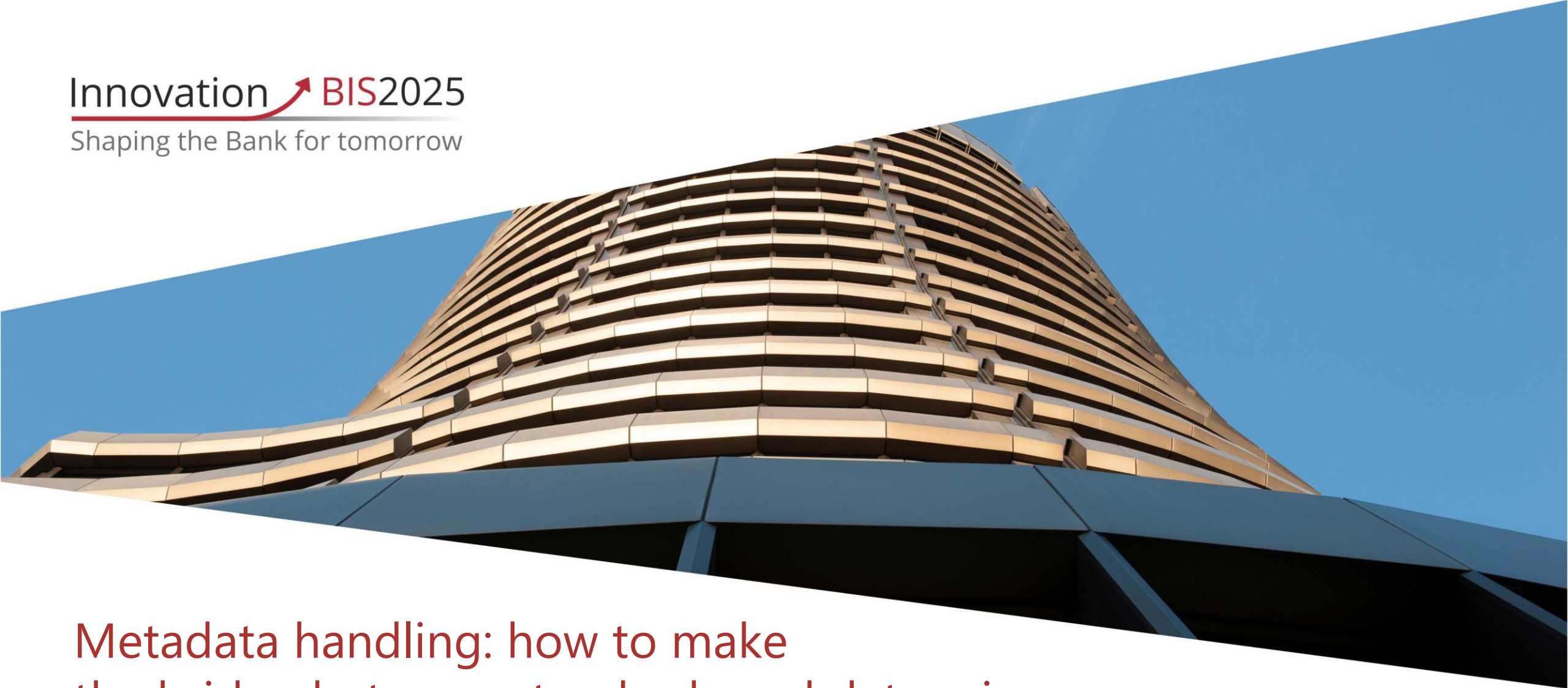


Innovation  BIS2025

Shaping the Bank for tomorrow



# Metadata handling: how to make the bridge between standards and data science

Olivier Sirello, Bank for International Settlements

HGL-MOS Workshop on the Modernization of Official Statistics

Geneva, 22 November 2023

The views expressed are those of the author and do not necessarily reflect those of the Bank for International Settlements. All errors are my own.

## Metadata: why are they relevant?

- Metadata are data about data
- Metadata are growing important in official statistics and carry huge potential:
  - Lineage and lifecycle
  - Data editing tasks and data quality
  - Event-driven metadata automations
  - Auto-tagging classifications and next generation of Data Catalog
  - Automate descriptions for data assets (e.g. GPT-based)
- However, metadata handling is often very challenging:
  - High multi-dimensionality
  - No specific workspace / "metadata" objects for many open-source languages

## Metadata handling: what are challenges?

- Metadata handling often faces three challenges:
  - Representation or multi-dimensionality
  - Heterogeneity of formats
  - Data and metadata relationships

## Metadata handling challenges

- **Representation or multi-dimensionality**

- Standard layout of data and metadata objects is tabular-based
  - Success of relational databases and DataFrames in several data science applications and across a large spectrum of languages (R – data.frame/tibble, Python – pd.DataFrame, Julia – DataFrames.jl)
- However, the high-dimensionality of metadata may clash with tabular layouts

Dimension	Value	Data lineage metadata
Country A	0	{...}
Country B	1	{...}





Dimension x Value	Formula 1	Steps	Log
Country A x 0	Mean()	{...}	{...}
Country A x 0	Median()	{...}	{...}
Country A x 0	Log()	{...}	{...}

## Metadata handling challenges

- **Heterogeneity of formats**

- Data assets are highly heterogenous, so are metadata
  - Text, documents, emails
  - Audio
  - Images and other visual contents

Dimension	Value	Compilation metadata
Country A	0	 
Country B	1	

## Metadata handling challenges

### ● Relationships

- Metadata have relationships *within* and *across* data structures

Dimension	Value A	Metadata 1	Metadata 2
Country A	0	Country A belongs to the European continent	Value is confidential
Country B		Country B belongs to the Asian continent	Value is unrestricted

Dimension	Value B
Country A	0
Country B	1

## Metadata handling: several solutions exist

- Going beyond standard tabular layouts:
  - Leverage more complicated objects (supporting n dimensionality)
    - Xarray or multi-indexing in Pandas (eg for Python)
- Manage relationships:
  - Use a standard
    - DDI or SDMX (eg attributes at different attachment levels)
- But several trade-offs: user-friendliness vs complexity, native solutions vs in-house
- **What is the experience of NSOs and CBs compiling official statistics while it comes to the handling of metadata, both for macro and micro data?**