SDMX 3.0 for Microdata: A Novel Preliminary Attempt for Retail Trade Index

Alicia Nieto Ramos, *Co-Head of Unit of Metadata.*

S.G. for Methodology and Sampling Design of Statistics Spain

SDMX Global Conference 2025, Rome

INE
Instituto Nacional de Estadística

Index

1. Introduction
2. First steps: Learning the model
   2.1. Analysis of the DSDs of PPP SDMX 2.1 microdata
   2.2. Analysis of RTI for aggregates used by Statistics Spain in SDMX 2.1
3. Implementing the PoC: DSD design for RTI with SDMX 3.0
   3.1. Description of our current data and metadata RTI structure
   3.2. Modeling the information: Selection of measures, attributes, and dimensions
   3.3. SDMX 3.0 Modeling Summary
4. Conclusions

1. Introduction

Providing some context...

- In 2022, Statistics Spain expressed a desire to analyze metadata standards in order to parallelize standardization and harmonization.
- The exploration of metadata standards at the microdata level began.
- A proof of concept was carried out in DDI Lifecycle 3.0 with microdata files from the Retail Trade Index survey.
- Once this was completed, it was decided to explore SDMX 3.0 for microdata, also working with microdata files from the Retail Trade Index survey.
- This proof of concept is what we are going to show in this presentation.

2. First steps: Learning the model

The first challenge that we had to face was the lack of publicly available examples of SDMX 3.0 metadata for microdata. To address this, we examined other SDMX metadata files for guidance, each of which served a specific purpose:

Example name	Is it for microdata?	Is it SDMX 3.0?	Is it Retail Trade Index?
DSDs of PPP SDMX 2.1 microdata	Yes ✓	No, ✗ it is SDMX 2.1.	No ✗
DSD design for RTI with SDMX 2.1	No ✗	No, ✗ it is SDMX 2.1.	Yes ✓

2.1. Analysis of the DSDs of PPP SDMX 2.1 microdata

We found the DSDs for PPP (Purchasing Power Parity) in SDMX 2.1 for microdata. Since both PPP and the Retail Trade Index are official statistical instruments (not private surveys) and their goal is to measure and compare relevant economic aspects, we decided to analyze the first one in depth.

Conclusions drawn from analyzing the PPP DSDs for microdata in SDMX 2.1:

- Learned the model: how concepts are defined (dimensions, attributes, measures, etc.).
- Understood how many DSDs need to be created per statistical process and why.
- Observed that in this PPP example, there are **only two microdata variables** (price and quantity), highlighting the potential complexity when dealing with larger datasets, as planned for the Proof of Concept with RTI.
- Saw how multiple measures can be included in a single DSD using SDMX 2.1.

2.2. Analysis of RTI for aggregates used by Statistics Spain in SDMX 2.1

We analyzed the DSDs of BCS (Business and Consumer Services survey) obtained from the Euro SDMX Registry to see which Codelists (used to metadata the RTI aggregates in SDMX 2.1) we would need to use when preparing the DSDs in SDMX 3.0 for microdata, so that they would remain consistent throughout the statistical production process.

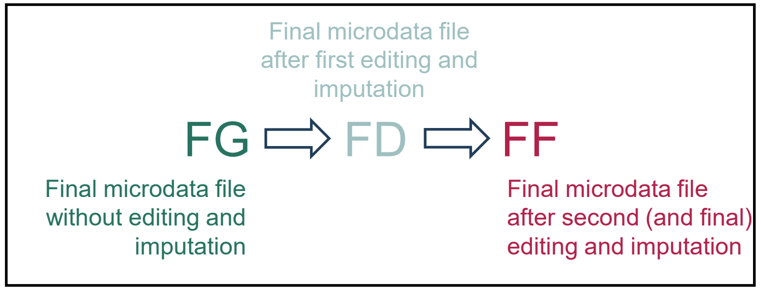
Finally, we saw that we would need to reuse:

1. **CL_ACTIVITY_BCS** (Required for SDMX 3.0 microdata)
   - Contains the economic activity codes (NACE) used in the flow.
   - Example: _T = Total activities, A01 = Agriculture, etc.
   - Defines the activity dimension in the data.
2. **CL_PRODUCT_BCS** (Required for SDMX 3.0 microdata)
   - Classification of products or specific subsectors according to the BCS flow.
   - Allows data to be broken down by product type within each activity.

3. Implementing the PoC: DSD design for RTI with SDMX 3.0

3.1. Description of our current data and metadata RTI structure

- A custom XML schema defines the metadata that governs its ETL process. This schema unifies all structural metadata for microdata files across production phases.
- For each time period, it has microdata in key-value and tabular format for the various production phases (FG, FD, FF), along with a unique XML variable-level metadata file designed based on auxiliary metadata variables called **qualifiers**.

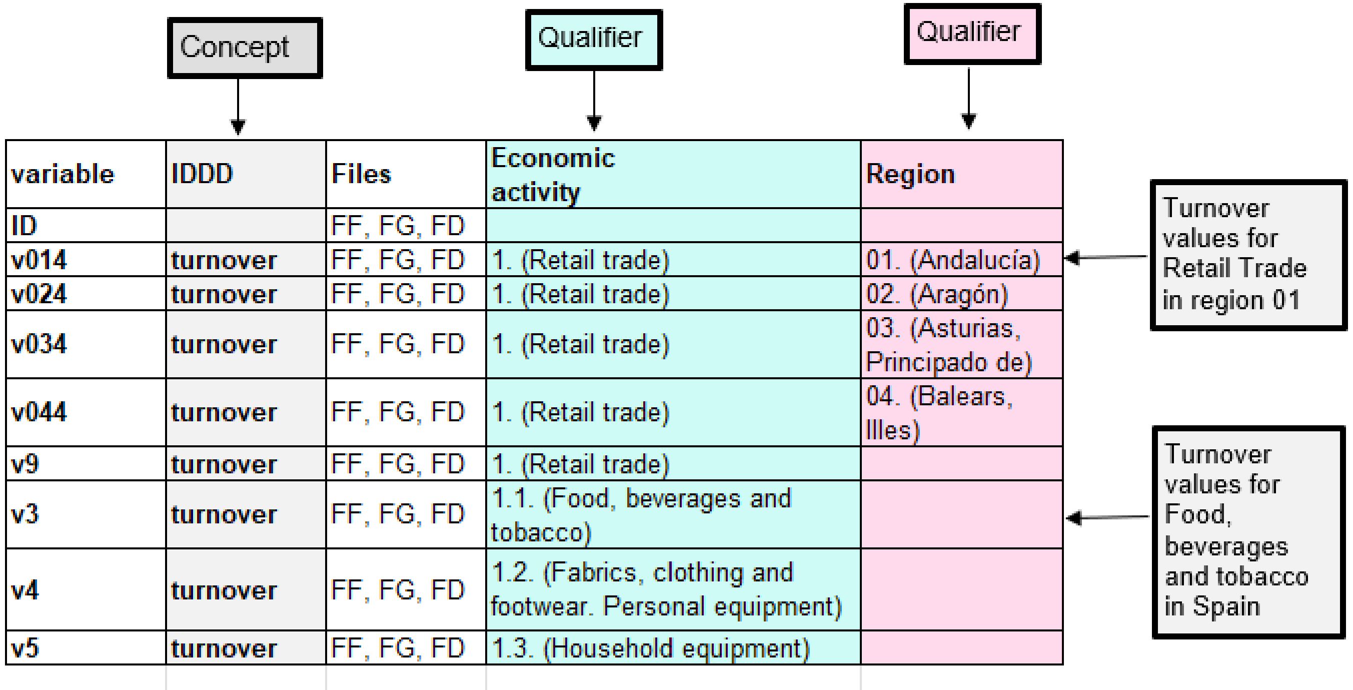


Final microdata file after first editing and imputation

FG ⇨ FD ⇨ FF

Final microdata file without editing and imputation

Final microdata file after second (and final) editing and imputation



Concept → IDDD

Qualifier → Economic activity

Qualifier → Region

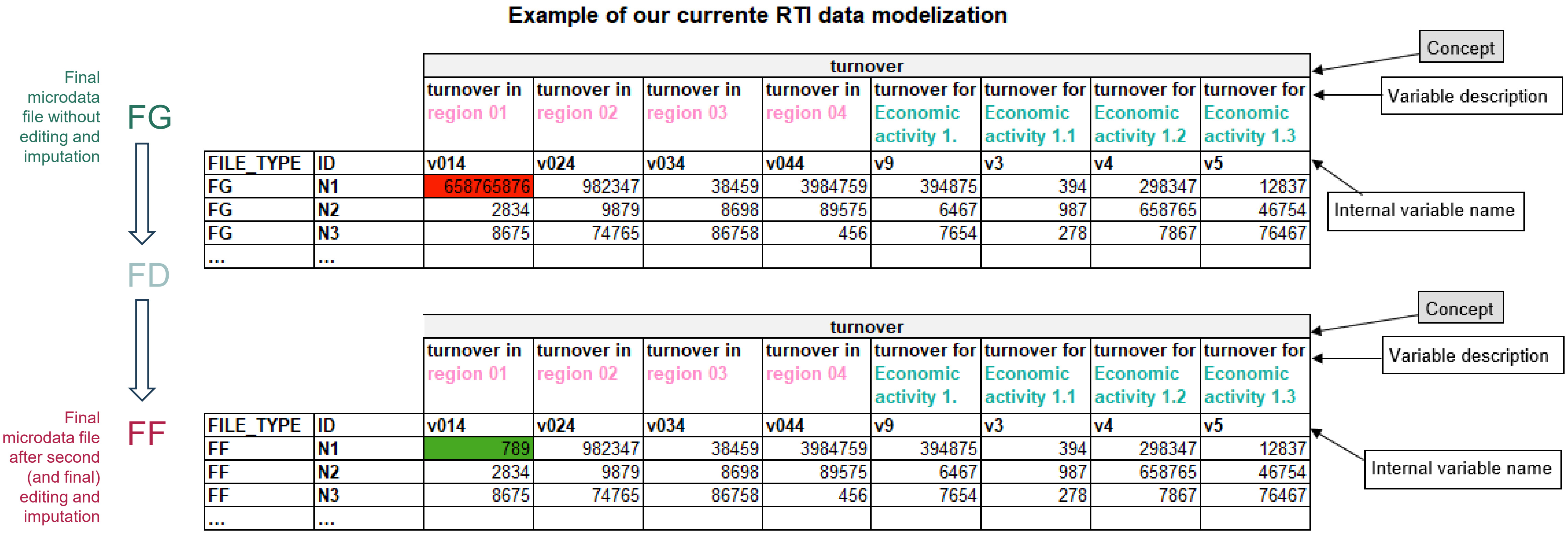
variable	IDDD	Files	Economic activity	Region
ID		FF, FG, FD		
v014	turnover	FF, FG, FD	1. (Retail trade)	01. (Andalucía)
v024	turnover	FF, FG, FD	1. (Retail trade)	02. (Aragón)
v034	turnover	FF, FG, FD	1. (Retail trade)	03. (Asturias, Principado de)
v044	turnover	FF, FG, FD	1. (Retail trade)	04. (Balears, Illes)
v9	turnover	FF, FG, FD	1. (Retail trade)	
v3	turnover	FF, FG, FD	1.1. (Food, beverages and tobacco)	
v4	turnover	FF, FG, FD	1.2. (Fabrics, clothing and footwear. Personal equipment)	
v5	turnover	FF, FG, FD	1.3. (Household equipment)	

Turnover values for Retail Trade in region 01

Turnover values for Food, beverages and tobacco in Spain

Image: Extract from the current RTI modeling based on qualifiers

(cont) 3.1. Description of our current data and metadata RTI structure



Example of our currente RTI data modelization

Final microdata file without editing and imputation
FG
FD
Final microdata file after second (and final) editing and imputation
FF
Concept
Variable description
Internal variable name
Concept
Variable description
Internal variable name

		turnover							
		turnover in region 01	turnover in region 02	turnover in region 03	turnover in region 04	turnover for Economic activity 1.	turnover for Economic activity 1.1	turnover for Economic activity 1.2	turnover for Economic activity 1.3
FILE_TYPE	ID	v014	v024	v034	v044	v9	v3	v4	v5
FG	N1	658765876	982347	38459	3984759	394875	394	298347	12837
FG	N2	2834	9879	8698	89575	6467	987	658765	46754
FG	N3	8675	74765	86758	456	7654	278	7867	76467
...	...								

		turnover							
		turnover in region 01	turnover in region 02	turnover in region 03	turnover in region 04	turnover for Economic activity 1.	turnover for Economic activity 1.1	turnover for Economic activity 1.2	turnover for Economic activity 1.3
FILE_TYPE	ID	v014	v024	v034	v044	v9	v3	v4	v5
FF	N1	789	982347	38459	3984759	394875	394	298347	12837
FF	N2	2834	9879	8698	89575	6467	987	658765	46754
FF	N3	8675	74765	86758	456	7654	278	7867	76467
...	...								

Image: extract from the FG and FF microdata files for the RTI, using the current modeling based on qualifiers.

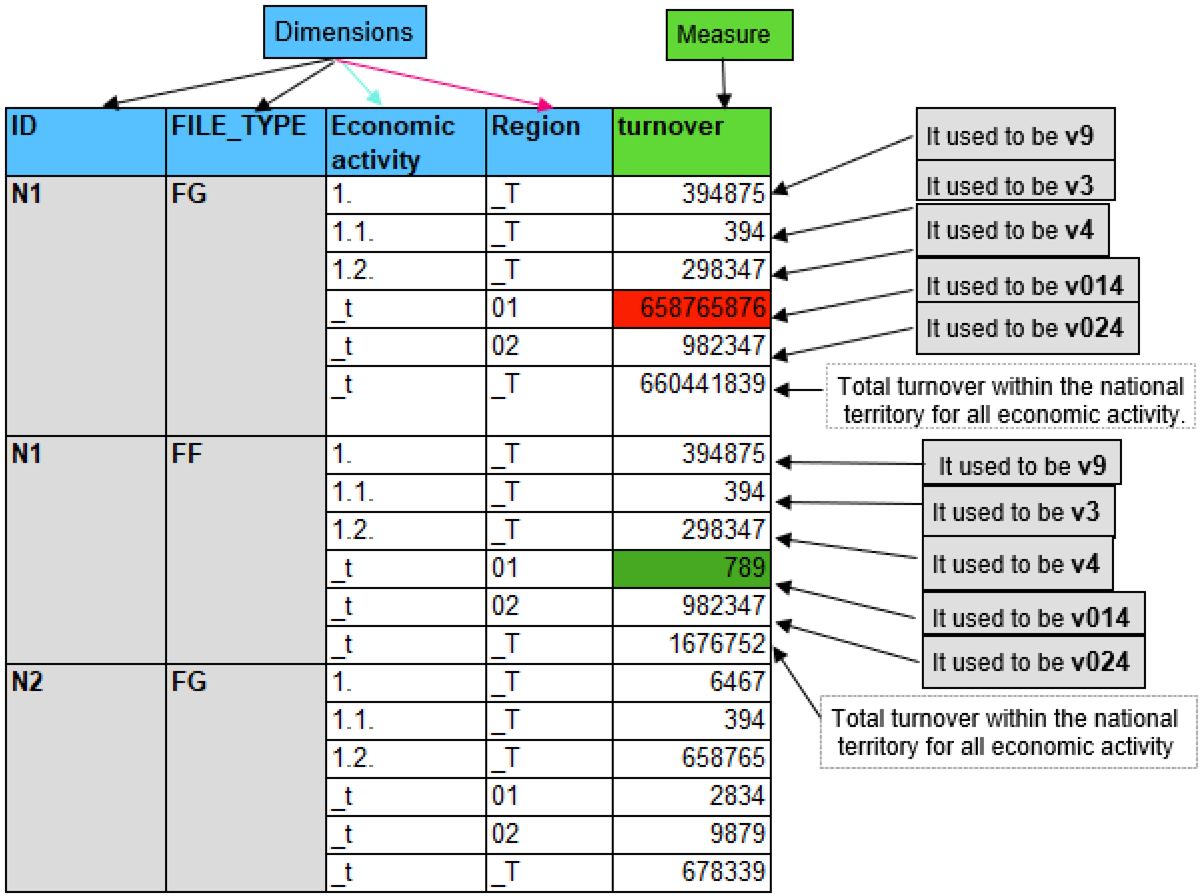
3.2. Modeling the information: Selection of measures, attributes, and dimensions

At the beginning of the proof of concept, we performed the modeling by defining as many measures as there were variables in our information system.

This resulted in an excessive number of measures, many of which captured data for the same concept (for example, 25 different measures for 'turnover’).

Therefore, we redesigned the model by restructuring the n-cube, adding new dimensions and reducing the number of measures:

Example of SDMX 3.0 modelization


Dimensions
Measure
It used to be v9
It used to be v3
It used to be v4
It used to be v014
It used to be v024
Total turnover within the national territory for all economic activity.
It used to be v9
It used to be v3
It used to be v4
It used to be v014
It used to be v024
Total turnover within the national territory for all economic activity

ID	FILE_TYPE	Economic activity	Region	turnover
N1	FG	1.	_T	394875
		1.1.	_T	394
		1.2.	_T	298347
		_t	01	658765876
		_t	02	982347
		_t	_T	660441839
N1	FF	1.	_T	394875
		1.1.	_T	394
		1.2.	_T	298347
		_t	01	789
		_t	02	982347
		_t	_T	1676752
N2	FG	1.	_T	6467
		1.1.	_T	394
		1.2.	_T	658765
		_t	01	2834
		_t	02	9879
		_t	_T	678339

Image. Extract from the new modeling: Economic activity and Region as dimensions, and a single measure: 'turnover'

3.3. SDMX 3.0 Modeling Summary

Dimensions:

Dimension List	CodeList
ID	
REGION	CL_REGIONAL_DEM_NUTS2024
ECONOMIC_ACTIVITY	CL_ACTIVITY_BCS
TIME_PERIOD	
FILE_TYPE	

Measures:

Measure List
ID_COLLECTION
SALES_PREMISES
TOTAL_SALES_PREMISES
LARGE_RETAIL_OUTLETS
TRADING_DAYS
HOLIDAYS_TRADING_DAYS
HAVE_STOCK
ECOMMERCE_TURNOVER
LRO_TURNOVER
TURNOVER
STOCK
EMPLOYEES
PAID_EMPLOYEES
NOTPAID_EMPLOYEES
LRO_PAID_EMPLOYEES
LRO_NOTPAID_EMPLOYEES
FLOOR_AREA_PREMISES

Attributes:

Attribute	CodeList	Format	Dimension/ Measure	Associated entity	OB/OP	Array
LARGE_SURFACES	{0,1}		DIMENSION	ID	OP	
TYPE_OF_CHAIN	{pcade,gcade,unil}		DIMENSION	ID	OP	
STOCK	{0,1}		DIMENSION	ID	OP	
STOCK_DUMMY	{0,1}		DIMENSION	ID	OP	
IT_IS_GAS	{0,1}		DIMENSION	ID	OP	
IMPUTATION_STATUS			DIMENSION	ID	OP	minOccurs=0 maxOccurs=5
WEEKLY	{0,1}		DIMENSION	ID	OP	
YEAR_ENTRY_IN_SAMPLE		AAAA	DIMENSION	ID	OP	
INCIDENCE	CL_INCIDENCE		DIMENSION	ID	OP	

4. Conclusions

- Great difficulty in **finding examples** of metadata in SDMX 3.0 for microdata.

- The information modeling we developed was **based** directly on the design of the stakeholder's **questionnaire**, which led to numerous semantic challenges when creating the new model:

  Example 1: In the RTI questionnaires, companies are asked to provide their own ID, which coexists with the ID from the statistical frame. *How should this be modeled?*
  Example 2: The questionnaire asks for the national total turnover, while at the same time the national total is calculated as the sum of all regions. *How can this be modeled to avoid having two combinations with the same dimensions producing a single measure (which would cause an error)?*

- Deciding what should be modeled as a measure, what as an attribute, and what as a dimension was a long and complex task — and one still open to improvement.

Thank you!

Contact us:

✉ alicia.nieto.ramos@ine.es

in http://linkedin.com/in/alicia-nieto-ramos/

✉ carmenelena.guaza.picallo@ine.es

✉ clara.marin.cuadros@ine.es

✉ metodologías@ine.es