SDMX GC 2023 – Advanced Capacity Building
SDMX, micro data in action

Stratos Nikoloutsos, Olivier Sirello (Bank for International Settlements)
A standard for micro data: trade-offs and challenges

1. **Standardisation but also customisation**
of data and metadata is key to facilitate interpretability, comparability and data lineage

2. **Reconciling micro and macro data: zooming in and out**
with the help of SDMX 3.0

3. **Proper modelling: from a top-down to a bottom-up approach**
to ensure consistency and standardisation across different data sets

4. **Data sharing with more performant and tailored queries**
made simpler and more efficient via SDMX
SDMX comes with a variety of tools

The **sdmx.io** ecosystem includes a collection of open source SDMX software tools cooperating to solve official statistics use cases.

[https://www.sdmx.io/tools/ecosystem/](https://www.sdmx.io/tools/ecosystem/)

SDMX/FMR is managed by the BIS and will be used as an example in this presentation although other tools may also offer the same features.
Standardisation of data and metadata is key to facilitate interpretability, comparability and data lineage.
The right balance between standardisation and customisation
Harmonized values across data sets, but also customisable code lists

Customisation and extension of code lists
- Micro data often come with the need for customised and/or extended code lists
- This feature is key for flexible maintenance of codes, for instance during the collection and compilation phases
  - In a security-by-security data set, new ISIN codes can be appended after each data collection round

Structure and representations maps
- Data sets with micro data typically are split into multiple tables
- Structure maps are key to describe the relationships between each of them
  - Structure maps can be used to describe the relationships between the columns of multiple tables (also allow to map custom internal codes to standard codes leveraging representation maps)

Structural metadata
- Structural metadata are key to describe statistical data, for each at data set, series, observation and measure level
- With SDMX 3.0, a list of values for attributes is allowed, increasing the flexibility of the data modelling notably required for micro data
  - In a security-by-security data set, SDMX 3.0 allows to set attributes for multiple measures, such as face, nominal and market value per each security per period
Extension of Codelists

ISO_3166-A2 (249 Codes)

How can I...
- add 10 legacy Country Codes
- add continents and regions (eg 29 Codes)
- change a few labels (eg 20 Codes)

Before SDMX 3.0:
- Create a new Country Codelist with 259 Codes
- Create a new Area Codelist with 278 Codes
- Create a new Country Codelist with 249 Codes
Extension of Codelists

ISO_3166-A2
(249 Codes)

How can I...
- add 10 legacy Country Codes
- add continents and regions (eg 29 Codes)
- change a few labels (eg 20 Codes)

With SDMX 3.0:
- Extend ISO_3166-A2 with 10 Codes
- Extend ISO_3166-A2 with 29 Codes
- Extend ISO_3166-A2 with 20 Codes

CL_COUNTRY
(10 Codes)

CL_AREA
(29 Codes)

CL_COUNTRY
(20 Codes)
### Arrays

**Observation Status**

<table>
<thead>
<tr>
<th>Position</th>
<th>Id</th>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
<td>Normal value</td>
<td>To be used as default value if no value is provided or when no special coded qualification is available.</td>
</tr>
<tr>
<td>2</td>
<td>B</td>
<td>Time series break</td>
<td>Observations are characterised as such when different content exists or a different methodology is applied.</td>
</tr>
<tr>
<td>3</td>
<td>D</td>
<td>Definition differs</td>
<td>Used to indicate slight deviations from the established methodology (footnote-type info).</td>
</tr>
<tr>
<td>4</td>
<td>E</td>
<td>Estimated value</td>
<td>Observation obtained through an estimation methodology (e.g. to produce back-casts).</td>
</tr>
<tr>
<td>5</td>
<td>F</td>
<td>Forecast value</td>
<td>Value deemed to assess the magnitude which a quantity will assume at some future point of time.</td>
</tr>
<tr>
<td>6</td>
<td>G</td>
<td>Experimental value</td>
<td>Data collected on the basis of definitions or (alternative) collection methods under development.</td>
</tr>
<tr>
<td>7</td>
<td>I</td>
<td>Value imputed by a receiving agency</td>
<td>Observation imputed by a receiving agency to replace or fill gaps in reported data series.</td>
</tr>
<tr>
<td>8</td>
<td>K</td>
<td>Data included in another category</td>
<td>This code is used when data for a given category are missing and are included in another category.</td>
</tr>
<tr>
<td>9</td>
<td>W</td>
<td>Includes data from another category</td>
<td>This code is used when data include another category, or go beyond the scope of the designated category.</td>
</tr>
<tr>
<td>10</td>
<td>O</td>
<td>Missing value</td>
<td>This code is to be used when no breakdown is made between the reasons why data are missing.</td>
</tr>
<tr>
<td>11</td>
<td>M</td>
<td>Missing value; data cannot exist</td>
<td>Used to denote empty cells resulting from the impossibility to collect a statistical value.</td>
</tr>
<tr>
<td>12</td>
<td>P</td>
<td>Provisional value</td>
<td>An observation is characterised as “provisional” when the source agency - while it bases its decisions on the best available data...</td>
</tr>
<tr>
<td>13</td>
<td>S</td>
<td>Strike and other special events</td>
<td>Special circumstances (e.g. strike) affecting the observation or causing a missing value.</td>
</tr>
<tr>
<td>14</td>
<td>L</td>
<td>Missing value; data exist but were not collected</td>
<td>Used, for example, when some data are not reported/disseminated because they are believed to be inappropriate or...</td>
</tr>
<tr>
<td>15</td>
<td>H</td>
<td>Missing value; holiday or weekend</td>
<td>Used in some holiday data flows.</td>
</tr>
<tr>
<td>16</td>
<td>Q</td>
<td>Missing value; suppressed</td>
<td>Used, for example, when data are suppressed due to statistical confidentiality considerations.</td>
</tr>
<tr>
<td>17</td>
<td>J</td>
<td>Derogation</td>
<td>Clause in an agreement (e.g. legal act, gentlemen’s agreement) stating that some provision is not applicable.</td>
</tr>
<tr>
<td>18</td>
<td>N</td>
<td>Not significant</td>
<td>Used to indicate a value which is not a “real” zero (e.g. a result of 0.0004 rounded to zero).</td>
</tr>
<tr>
<td>19</td>
<td>U</td>
<td>Low reliability</td>
<td>This indicates existing observations, but for which the user should also be aware of the...</td>
</tr>
</tbody>
</table>
Arrays as value for Attribute/Measure

How can I...
- provide 2 statuses for an observation?
- provide more than one statuses without knowing the exact number on design time?

Before SDMX 3.0:
- Add two OBS_STATUS Attributes
- ?

OBS_STATUS_1: F (Forecast)
OBS_STATUS_2: U (Low reliability)

OBS_STATUS_N: D (Definition differs)
Arrays as value for Attribute/Measure

**Observation Status**

*How can I...*
- provide 2 statuses for an observation?
- provide more than one statuses without knowing the exact number on design time?

*With SDMX 3.0:*
- Add one OBS_STATUS Attribute with max = 2
- Add one OBS_STATUS Attribute with unbounded upper limit (or high enough)

**OBS_STATUS:**
- [F (Forecast), U (Low reliability)]
- [F (Forecast), U (Low reliability), ... D (Definition differs)]
Easy reconciliation of micro and macro data
with the help of SDMX

Hierarchies to quickly drill down from aggregates or conversely

Mappings to link representations and understand relationships between concepts

Track the validity of concepts and codes, their groupings, as well as their mappings, across time

Attributes at multiple levels with fine grained attachment
From micro to macro and from macro to micro
Quickly drill down from aggregates and conversely

Hierarchies to zoom in and zoom out
- Hierarchies are key to *drill down on the most granular level* from aggregates
- Groups and hierarchies share the same standardized codes, ensuring consistency
  - Share the codes across different groups, such as a country belonging to multiple economic groupings
  - Derive from the hierarchy the underlying entities that have been aggregated

Mappings to better understand relationships between concepts
- Map representations, also leveraging regular expressions, to other representations and concepts
- Mapping also includes free text and can be *one-to-many or many-to-many*
  - Map the initial two letters of the ISIN code to the country dimension

Attributes and multiple measures
- Measure-specific attributes: an attribute can be explicitly related to one or more measures
  - As an example, it might be possible define attribute A “1” for Gender and attribute A “2” for Occupation and filter according to their values
Hierarchies in SDMX 3.0

Viewing: Country groupings in the Data Dictionary [1.0]

- G000 - World
  - G100 - Africa
    - G110 - Northern Africa
      - DZ - Algeria
      - EG - Egypt
      - LY - Libya
      - MA - Morocco
      - SD - Sudan
      - TN - Tunisia
      - EH - Western Sahara
    - G120 - Sub-Saharan Africa
      - G121 - Eastern Africa
        - IO - British Indian Ocean Territory (the)
        - BI - Burundi
        - KM - Comoros
        - DJ - Djibouti
        - ER - Eritrea

Viewing: Country groupings in the Data Dictionary [1.0]

- EU - European Union (the)
  - BE - Belgium
  - DE - Germany
  - DK - Denmark
  - ES - Spain
  - FR - France
  - GB - United Kingdom (the)
  - GR - Greece
  - IE - Ireland
  - IT - Italy
  - LU - Luxembourg
  - NL - Netherlands (the)
  - PT - Portugal

- EU- - European Union (the) excl. Luxembourg
- XM - Euro Area
- EURO- - Euro Area excl. Luxembourg

Time period: 1980
Hierarchies in SDMX 3.0

May be related to a context (Hierarchy Association)

Linked to an object (eg a Dimension)

For a given context (eg a Dataflow)
Structure and Representation Mappings

A → ANNUAL
EL → GREECE
GR → GREECE

ANNUAL
GREECE
2009
(GDP/C) 29828.76
(LFORCE) 67.51
USD
PC
K

A
EL
GDP/C
2009
29828.76
USD
K

ANNUAL
GR
LFORCE
2009
67.51
PC
Structure and Representation Mappings

A
EL
GDP/C
2009
29828.76
US dollars
K

ANNUAL
GREECE
2009
(GDP/C) 29828.76
(LFORCE) 67.51
USD
PC
K

US Dollars → USD
Structure and Representation Mappings

Annual → A
Annually → A
Yearly → A

CH1234567890 → CH

An International Securities Identification Number (ISIN) is a code that uniquely identifies a security globally for the purposes of facilitating clearing, reporting and settlement of trades. (ISO 6166)

^CH → CH
^([A-Z]{2}) → \1
Proper modelling
of micro data ensures consistency and standardisation across different data sets

Leveraging patterns to define the uncoded dimensions of the data

ValueLists as flexible sets of characters

Multiple measures to model micro data as one observation

Attributes can take unbounded lists of values to give more flexibility in the modelling
Achieve a flexible modeling, a bottom-up perspective

Leveraging patterns to define the uncoded dimensions of the data

A component based on a value domain that follows a pattern, without requiring the creation of a list of code.

It is possible to derive the country ISO2 code from the first two letters of the column “ISIN code” from a security-by-security database.

ISIN code -> Reference area, thus CH0000000000 -> CH

Multiple measures to model micro data as one observation

More than one measurement per record, allowing also to provide fine grained metadata per measure – rationalizing/simplifying data modelling.

A security-by-security table may contain three measures for the amount outstanding, face, nominal and market value.

SDMX allows to define attributes at the measure level, for example to flag confidential only some specific values.

It also allows to define several statuses for a given value, eg provisional and unvalidated value for market value on 2023-20.

<table>
<thead>
<tr>
<th>Time period</th>
<th>ISIN code</th>
<th>Face value</th>
<th>Nominal value</th>
<th>Market value</th>
</tr>
</thead>
<tbody>
<tr>
<td>2023-10</td>
<td>CH0123456789</td>
<td>12</td>
<td>11.5</td>
<td>14&lt;sub&gt;P,V&lt;/sub&gt;</td>
</tr>
<tr>
<td>2023-09</td>
<td>CH0123456789</td>
<td>12</td>
<td>11.6&lt;sub&gt;CONF&lt;/sub&gt;</td>
<td>13</td>
</tr>
<tr>
<td>2023-08</td>
<td>CH0123456789</td>
<td>12</td>
<td>12</td>
<td>15</td>
</tr>
</tbody>
</table>
Achieve a flexible modeling, a bottom-up perspective

Leveraging patterns to define the uncoded dimensions of the data

A component based on a value domain that follows a pattern, without requiring the creation of a list of code

It is possible to derive the country ISO2 code from the first two letters of the column “ISIN code” from a security-by-security db ISIN code → Reference area, thus CH0000000000 -> CH

```
<str:Dimension>
  <str:ConceptIdentity>urn:...Concept=SDMX:CONCEPTS(1.0).ISIN</str:ConceptIdentity>
  <str:LocalRepresentation>
    <str:TextFormat pattern="^[A-Z]{2}[A-Z0-9]{9}\d$" />
  </str:LocalRepresentation>
</str:Dimension>
```

Two-letter country code: exactly two uppercase letters from 'A' to 'Z'
The 9-character alphanumeric code that follows the country code: can be uppercase letters or digits from '0' to '9'.
Achieve a flexible modeling, a bottom-up perspective

Leveraging patterns to define the uncoded dimensions of the data

A component based on a value domain that follows a pattern, without requiring the creation of a list of code

It is possible to derive the country ISO2 code from the first two letters of the column “ISIN code” from a security-by-security db

ISIN code -> Reference area, thus CH0000000000 -> CH
Achieve a flexible modeling, a bottom-up perspective

Multiple measures to model micro data as one observation

More than one measurement per record, allowing also to provide fine grained metadata per measure – rationalizing/simplifying data modelling

A security-by-security table may contain three measures for the amount outstanding, face, nominal and market value. SDMX allows to define attributes at the measure level, for example to flag confidential only some specific values. It also allows to define several statuses for a given value, eg provisional and unvalidated value for market value on 2023-20.

<table>
<thead>
<tr>
<th>Time period</th>
<th>ISIN code</th>
<th>Face value</th>
<th>Nominal value</th>
<th>Market value</th>
</tr>
</thead>
<tbody>
<tr>
<td>2023-10</td>
<td>CH0123456789</td>
<td>12</td>
<td>11.5</td>
<td>14&lt;sup&gt;P,V&lt;/sup&gt;</td>
</tr>
<tr>
<td>2023-09</td>
<td>CH0123456789</td>
<td>12</td>
<td>11.6&lt;sup&gt;CONF&lt;/sup&gt;</td>
<td>13</td>
</tr>
<tr>
<td>2023-08</td>
<td>CH0123456789</td>
<td>12</td>
<td>12</td>
<td>15</td>
</tr>
</tbody>
</table>
Achieve a flexible modeling, a bottom-up perspective

Multiple measures to model micro data as one observation

More than one measurement per record, allowing also to provide fine grained metadata per measure – rationalizing/simplifying data modelling

A security-by-security table may contain three measures for the amount outstanding, face, nominal and market value. SDMX allows to define attributes at the measure level, for example to flag confidential only some specific values. It also allows to define several statuses for a given value, e.g. provisional and unvalidated value for market value on 2023-20

<table>
<thead>
<tr>
<th>Time period</th>
<th>ISIN code</th>
<th>Face value</th>
<th>Nominal value</th>
<th>Market value</th>
</tr>
</thead>
<tbody>
<tr>
<td>2023-10</td>
<td>CH0123456789</td>
<td>12</td>
<td>11.5</td>
<td>14&lt;sup&gt;P&lt;/sup&gt;, 14&lt;sup&gt;V&lt;/sup&gt;</td>
</tr>
<tr>
<td>2023-09</td>
<td>CH0123456789</td>
<td>12</td>
<td>11.6&lt;sup&gt;CONF&lt;/sup&gt;</td>
<td>13</td>
</tr>
<tr>
<td>2023-08</td>
<td>CH0123456789</td>
<td>12</td>
<td>12</td>
<td>15</td>
</tr>
</tbody>
</table>
Achieve a flexible modeling, a bottom-up perspective

Multiple measures to model micro data as one observation

More than one measurement per record, allowing also to provide fine grained metadata per measure – rationalizing/simplifying data modelling

A security-by-security table may contain three measures for the amount outstanding, face, nominal and market value. SDMX allows to define attributes at the measure level, for example to flag confidential only some specific values.

It also allows to define several statuses for a given value, eg provisional and unvalidated value for market value on 2023-20

```xml
<str:Measure>
  <str:ConceptIdentity>urn:...Concept=SDMX:CONCEPTS(1.0).FACE</str:ConceptIdentity>
</str:Measure>
<str:Measure>
  <str:ConceptIdentity>urn:...Concept=SDMX:CONCEPTS(1.0).NOMINAL</str:ConceptIdentity>
</str:Measure>
<str:Measure>
  <str:ConceptIdentity>urn:...Concept=SDMX:CONCEPTS(1.0).MARKET</str:ConceptIdentity>
</str:Measure>
```
Achieve a flexible modeling, a bottom-up perspective

Multiple measures to model micro data as one observation

More than one measurement per record, allowing also to provide fine grained metadata per measure – rationalizing/simplifying data modelling

A security-by-security table may contain three measures for the amount outstanding, face, nominal and market value. SDMX allows to define attributes at the measure level, for example to flag confidential only some specific values. It also allows to define several statuses for a given value, eg provisional and unvalidated value for market value on 2023-20

```xml
<str:Attribute>
  <str:ConceptIdentity>urn:...Concept=SDMX:CONCEPTS(1.0).CONF</str:ConceptIdentity>
  <str:AttributeRelationship>
    <str:Observation/>
  </str:AttributeRelationship>
  <str:MeasureRelationship>
    <str:Measure>FACE</str:Measure>
  </str:MeasureRelationship>
</str:Attribute>
```
A security-by-security table may contain three measures for the amount outstanding, face, nominal and market value. SDMX allows to define attributes at the measure level, for example to flag confidential only some specific values. It also allows to define several statuses for a given value, eg provisional and unvalidated value for market value on 2023-2027.

```xml
<str:Attribute>
  <str:ConceptIdentity>urn:...Concept=SDMX:CONCEPTS(1.0).CONF</str:ConceptIdentity>
  <str:LocalRepresentation minOccurs="1" maxOccurs="3">
    <str:Enumeration>urn:...Codelist=SDMX:CL_OBS_STATUS(1.0)</str:Enumeration>
  </str:LocalRepresentation>
...
</str:Attribute>
```
Data sharing
made simpler and more efficient via SDMX

SDMX Open API to programmatically access data, metadata and structures

Customizable data queries, custom filters or specific values retrieval

A variety of formats to retrieve and store data and metadata

Interoperability with other formats such as xBRL
Open API, more formats and interoperability

A powerful new Open API with increased flexibility
- Accessing data, metadata, structures within a client application also to ease their maintenance
- New parameters and operators for more flexible data and metadata querying

Querying for data for a range of values of a measure or attribute, including string matching

A variety of formats to retrieve and store data and metadata
- More data and metadata formats
- Combining data with reference metadata

New powerful XML messages, covering all standards
JSON targeting data visualization with combined data and structural metadata
CSV for human readable/processible datasets

Interoperability with other formats (xBRL)
- Ongoing work to link the two standards

xBRL-SDMX converter to facilitate the interoperability
SDMX 3.0 RESTful API v2.0.0

- The normative part of the specification, i.e. the Open API definition
- The Developers' documentation, including a cheat sheet
- Request features and report issues on GitHub
The structure queries

- Re-organized and enriched
- Supports multiple instances of search terms, wildcarding

https://host/structure/type/agency/id/version/item?detail&references

The type of structure:
- datastructure, metadatastructure, categoriescheme, conceptscheme, codelist, hierarchy, hierarchyassociation, valuelist, agency schemes, dataproviderscheme, dataconsumerscheme, organisationunitscheme, dataflow, metadatamount, reportingtaxonomy, provisionagreement, structuremap, representationmap, conceptschememap, categorieschememap, organisationschememap, reportingtaxonomymap, process, categorisation, dataconstraint, metadataconstraint, structure, transformationscheme, rulesetscheme, userdefinedoperatorscheme, customtypescheme, namepersonalisationscheme, vtlmappingscheme

The unique identifier:
- Agency ID
- Artefact ID
- Version
- Item ID (for Item Schemes)

Amount of information:
- allstubs, referencestubs, allcompletestubs, referencecompletestubs, referencepartial, raw, partialraw, full

References to be returned:
- none, parents, parentsandsiblings, ancestors, children, descendants, all, a resource type
The data queries

https://host/data/context/agency/id/version/key?c

Key(s) of the series to be returned:
- eg M.GR.EUR.SP00
- with wildcarding:
  - eg M.*.EUR.SP00

With support for multiple keys:
- eg M.GR.EUR.SP00, M.CY.EUR.SP00

The unique identifier of the context:
- Agency ID
- Artefact ID
- Version

The context of data retrieval:
- datastructure, dataflow, provisionagreement

Component-based filters (for any Dimension, Attribute or Measure):
- eg c[REF_AREA]=CH&c[CONF_STATUS]=F

Support for operators:

Support for operators:
- eq Equal
- ne Not equal
- lt Less than
- le Less than or equal to
- gt Greater than
- ge Greater than or equal to
- co Contains
- nc Does not contain
- sw Starts with
- ew Ends with
- nd And
- or Or

UpdatedAfter
&firstNObservations
&lastNObservations
&attributes
&measures
&dimensionAtObservation
&includeHistory
The data queries

https://host/data/context/agency/id/version/key?c

- Retrieves what changed since supplied timestamp. Must be percent-encoded (e.g.: 2009-05-15T14%3A15%3A00%2B01%3A00)
- Maximum number of observations starting from the first observation
- Maximum number of observations counting back from the most recent observation
- Id of the dimension at the observation level

The attributes to be returned:
- dsd, msd, dataset, series, obs, all, none, {attribute_id}

The measures to be returned:
- all, none, {measure_id}

Whether to return vintages

&updatedAfter
&firstNObservations
&lastNObservations
&attributes
&measures
&dimensionAtObservation
&includeHistory
Other queries

- **Data validity**
  
  https://host/schema/context/agency/id/version?dimensionAtObservation&explicitMeasure

- **Data availability**
  
  https://host/availability/context/agency/id/version/key/componentId?c
  &updatedAfter
  &references
  &mode

- **Metadata**
  
  https://host/metadata/metadataset/provider/id/version?detail
  
  https://host/metadata/metadataflow/agency/id/version/provider?detail
  
  https://host/metadata/structure/type/agency/id/version/provider?detail
## The formats

<table>
<thead>
<tr>
<th>SDMX-ML Data</th>
<th>application/vnd.sdmx.data+xml;version=3.0.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>SDMX-ML Structure</td>
<td>application/vnd.sdmx.structure+xml;version=3.0.0</td>
</tr>
<tr>
<td>SDMX-ML Metadata</td>
<td>application/vnd.sdmx.metadata+xml;version=2.0.0</td>
</tr>
<tr>
<td>SDMX-JSON Data</td>
<td>application/vnd.sdmx.data+json;version=2.0.0</td>
</tr>
<tr>
<td>SDMX-JSON Structure</td>
<td>application/vnd.sdmx.structure+json;version=2.0.0</td>
</tr>
<tr>
<td>SDMX-JSON Metadata</td>
<td>application/vnd.sdmx.metadata+json;version=2.0.0</td>
</tr>
<tr>
<td>SDMX-CSV Data</td>
<td>application/vnd.sdmx.data+csv;version=1.0.0</td>
</tr>
<tr>
<td>SDMX-CSV Metadata</td>
<td>application/vnd.sdmx.metadata+csv;version=2.0.0</td>
</tr>
</tbody>
</table>
The SDMX v2 API in **action**

- See the API spec on [SwaggerHub](https://app.swaggerhub.com/home)
Thank you!

Olivier.Sirello@bis.org
Stratos.Nikoloutsos@bis.org