10TH EXPERT GROUP MEETING ON

Statistical Data and Metadata eXchange

JANUARY 25-28, 2021

SDMX modelling

January 26, 2021

Gyorgy Gyomai
OECD

sdmx
Statistical Data and Metadata eXchange

Outline

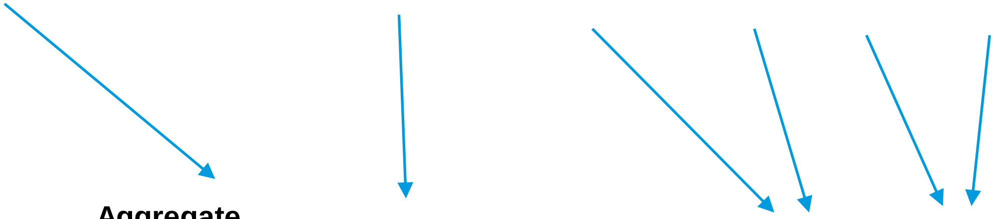
- Organising data models at a data-warehouse scale:
  - Make data maintainable and combinable
  - Connecting micro and macro
  - Modelling basic statistics and derived indicators together
- -------------------
- A deep-dive:
  - Modelling the Unit of Measure

Separate micro and aggregate data-models

Microdata Dataset

DIM	ATTR_C	ATTR_U	ATTR_C	MEAS	MEAS	ATTR_C	ATTR_C
Entity ID	**Country of residence**	**Address**	**Gender**	**Height**	**Weight**	**UoM Height**	**UoM Weight**
XY	AUS	…	F	150	50	cm	kg
WZ	AUS	…	M	170	80	cm	kg

DIM	Dimension
ATTR_C	Controlled attribute
ATTR_U	Uncontrolled attribute
MEAS	Measure



Aggregate Dataset

DIM	DIM	DIM	MEAS	ATTR_C
Country of residence	**Gender**	**Measure**	**Observation**	**Unit of Measure**
AUS	F	Average height	150	cm
AUS	M	Average height	170	cm
AUS	_T	Average height	160	cm
AUS	F	Average weight	50	kg
AUS	M	Average weight	80	kg
AUS	_T	Average weight	65	kg

sdmx

Statistical Data and Metadata eXchange

Sifting together micro and aggregate models

Joint Dataset

DIM	DIM	DIM	MEAS	ATTR_C	DIM	ATTR_U
Country of residence	**Gender**	**Measure**	**Observation**	**Unit of Measure**	**Entity ID**	**Address**
AUS	F	Average height	150	cm	_T	_Z
AUS	M	Average height	170	cm	_T	_Z
AUS	_T	Average height	160	cm	_T	_Z
AUS	F	Average weight	50	kg	_T	_Z
AUS	M	Average weight	80	kg	_T	_Z
AUS	_T	Average weight	65	kg	_T	_Z
AUS	F	Average height	150	cm	XY	…
AUS	F	Average weight	50	kg	XY	…
AUS	M	Average height	170	cm	WZ	…
AUS	M	Average weight	80	kg	WZ	…

Sifting together raw data and indicators

Aggregate data and indicators

DIM	DIM	DIM	MEAS	ATTR_C	
Country of residence	**Gender**	**Measure**	**Observation**	**Unit of Measure**	
AUS	F	Average height	150	cm	
AUS	M	Average height	170	cm	
AUS	_T	Average height	160	cm	
AUS	F	Average weight	50	kg	
AUS	M	Average weight	80	kg	
AUS	_T	Average weight	65	kg	
AUS	F	Average BMI	22.22	kg/m^2	
AUS	M	Average BMI	27.68	kg/m^2	
AUS	_T	Average BMI	24.95	kg/m^2	
AUS	_Z/_T	Gender Diff Height	-20	cm	*absolute
AUS	_Z/_T	Gender Diff Weight	30	kg	*absolute
AUS	_Z/_T	Gender Diff BMI	21.88	% (of national BMI)	*relative

The bigger picture

- In a large multi-domain data-warehouse how to co-ordinate concept schemes?
  - Our choice: the global DSD model (separate concept schemes for each domain, with concepts and code-lists reuse promoted)
  - For Referential metadata a shared MSD for the entire data-warehouse
- Multiple data spaces (collection, processing, dissemination, etc.) help organise functionally different shapes of the same data for dissemination, production and exchange, or switch from a provider specific model to and integrated/harmonised data model.
- The role of mappings and VTL
  - EDD
  - SDMX Structure Sets
  - VTL – transformation rules

Abstraction

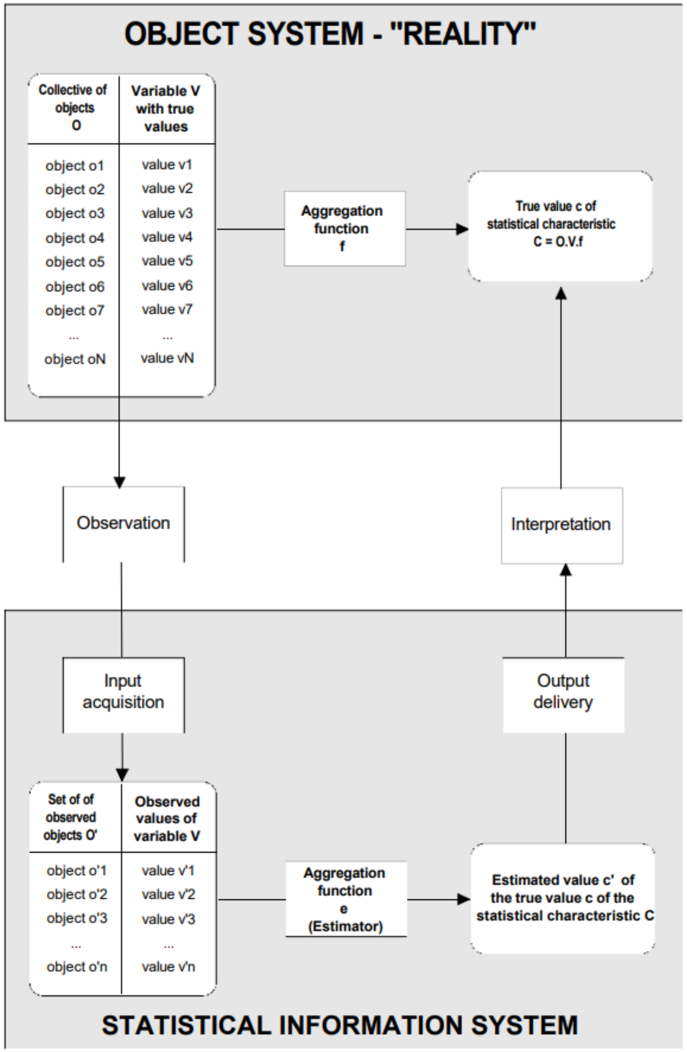
- In the UNECE 1995 *Guidelines for the modelling of statistical data and metadata*

  <Object, Variable, Aggregator function>

- A similar, alternative starting point

  Object.Property = Value [Unit of measure]

  (variable and aggregator function considered together in this alternative model)



OBJECT SYSTEM - "REALITY"
Aggregation function f
True value c of statistical characteristic C = O.V.f
Observation
Interpretation
Input acquisition
Output delivery
Aggregation function e (Estimator)
Estimated value c' of the true value c of the statistical characteristic C
STATISTICAL INFORMATION SYSTEM

Collective of objects O	Variable V with true values
object o1	value v1
object o2	value v2
object o3	value v3
object o4	value v4
object o5	value v5
object o6	value v6
object o7	value v7
...	...
object oN	value vN

Set of of observed objects O'	Observed values of variable V
object o'1	value v'1
object o'2	value v'2
object o'3	value v'3
...	...
object o'n	value v'n

The Mountain example

Object: Chimborazo

Property: height

Value: 6263

Unit of measure: meters

CC BY-SA: **David Torres Costales**

Statistical Data and Metadata eXchange

A National Accounts example

Object: USA; General (or central) government; Q3 2012

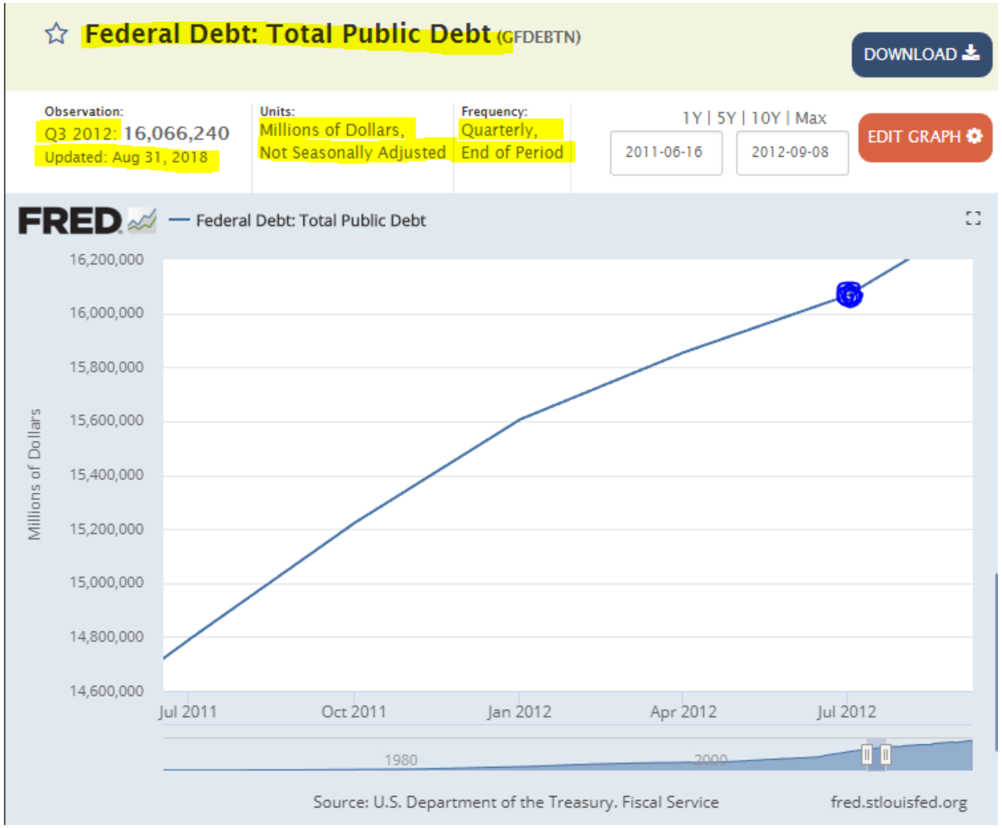
Property: Debt; Gross; Consolidated; End of period

Value: 16 066 240

Unit of measure: USD; millions;

Less obvious associations:

Quarterly, non-seasonally adjusted, current price, Aug 31 2018 vintage, level (as opposed to transformations)



☆ Federal Debt: Total Public Debt (GFDEBTN)

Observation:
Q3 2012: 16,066,240
Updated: Aug 31, 2018

Units:
Millions of Dollars,
Not Seasonally Adjusted

Frequency:
Quarterly,
End of Period

1Y | 5Y | 10Y | Max

2011-06-16 — 2012-09-08

Source: U.S. Department of the Treasury. Fiscal Service

fred.stlouisfed.org

Units of measure

Object: Chimborazo
Property: height
Value: 6263
Unit of measure: metres

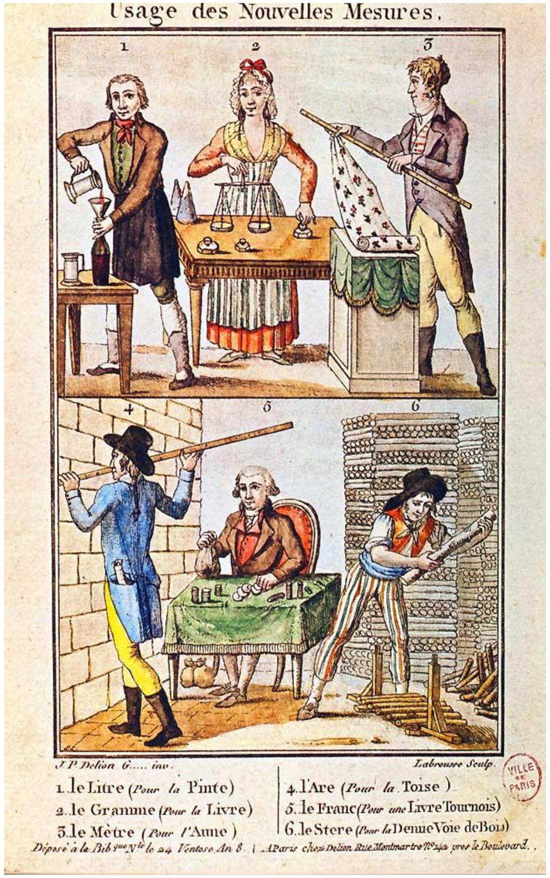
The ‘evolution’ of a metre:
1791: On ten millionth of the distance between the equator and the north pole, a meridian quadrant.
1927: Platinum-iridium bar at melting point of ice, atmospheric pressure, supported by two rollers
1983: Length of the path travelled by light in vacuum in 1/299 792 458 seconds

The main benefit of systematic use of units of measurement is the wide applicability of dimensional analysis:

- Two sides of an equation should share the same unit of measure
  e.g. x [metres] = y [metres]
- Addition/subtraction should only be done on quantities with shared units of measure
  e.g. x [metres] + y [metres] = z [metres]
- Derived indicators (through multiplication/division) should receive derived units according to the manipulations applied to the values themselves
  e.g. z [metres/seconds]=d [metres]/ t [seconds]
- Changing of unit of measure is straightforward
  a = average Earth-Sun distance: 1 [astronomical unit] = 149.6 *10^9 [metres]
  b = average Jupiter-Sun distance: 778.5*10^9 [metres] = 5.2 [astronomical units]

x [old units]= *x* (1/*n*) [*new* units]

where *n* is the size of the new unit in old units


Usage des Nouvelles Mesures.
1
2
3
4
5
6
J. P. Delion G..... inv.
Labrousse Sculp.
1. le Litre (Pour la Pinte)
2. le Gramme (Pour la Livre)
3. le Mètre (Pour l'Aune)
4. l'Are (Pour la Toise)
5. le Franc (Pour une Livre Tournois)
6. le Stere (Pour la Demie Voie de Bois)
Déposé à la Bibque Nle le 24 Ventose An 8. | A Paris chez Delion Rue Montmartre No 242 pres le Boulevard.
VILLE DE PARIS

Attributes or dimensions?

- The value of Unit of Measure is determined by multiple concepts: often property related, but also purely measurement standard related, and rarely object specific.

[Country:USA] [Transaction(Indicator):Debt]
[KeyUnitDeterminant(UnitOfMeasure):NationalCurrency] [UnitMultiplier:millions] [Adjustment:Nsa] [Prices:Current Price][Transformation: None]

→ 'current price USD, millions, non-seasonally adjusted'
→ or 'USD, millions' might be sufficient in certain contexts

- At the OECD we are component agnostic.

- Considerations for the A/D choice in general
  - What is the widest context in which the data will appear? Is the concept needed for disambiguation? E.g. for 'base year', in exchanges: attribute vs. in production dbase: dimension
  - Is the ability to query, ability to constrain needed?

- Annotations to control presentation
  - UNIT_MEASURE_CONCEPTS annotation to list the concepts that make up the unit of measure label
  - UNIT_MEASURE_LABEL annotation in combination with a custom built attribute used optionally when finer-grain editorial control is needed
  - Attachment level inferred; as the union of attachment-dimensions of attributes indicated by UNIT_MEASURE_CONCEPTS + the dimensions in it

Unit measure examples from NA CL_UNIT

- H1: Euro area-18 countries: FR,BE,LU,NL,DE,IT,IE,PT,ES,FI,AT,GR,SI,CY,EE,LV,MT,SK (how is this a unit of measure?)
- PN: Pure number (intriguing)
- RO: Ratio (of what, another mysterious unit of measure)
- TSO2E_R_POP: Tonnes of SO2-equivalent per capita (good)
- EUR_R_POP: Euro; ratio to total population (not so great)
- PU_R_POP: Per capita, US $, PPP converted (good)

Expressing monetary value and exchange rates

- SPL: Seborga, Luigini (I had to google this but it is as valid - if not more - than Bitcoins)
- XDC: Domestic currency (incl. conversion to current currency made using a fixed parity) [Context specific - but OK, if one pays attention, and substitutes the country in context - even dimensional calculus is possible]
- XDC_R_B1G: Domestic currency (incl. conversion to current currency made using a fix parity); ratio to gross value added (it would be simpler to say 'times' or 'percent of Gross Value Added', there are similar cases, e.g. Indices where the original unit of measurement matters.)
- XDC_R_B2G_S11: percentage of gross operating surplus of non-financial corporations (a similar measure, with a wording much closer to home, I'd only drop 'age' from percentage)
- XDO: Other currencies not included in the SDR basket, exc. gold and SDRs (it looks similar to XDC, but there is no context where this unit can meaningfully be associated with a number to represent a well defined quantity)
- XXEXE: Exchange rate (end of period): currency of area per currency of counterpart area (not ideal to have a context driven UoM, but if needed for brevity - or filtering/pivoting, it should not include 'Exchange rate' as that is really the property here)
- CD: National currency per US dollar (unit for exchange rates and PPP) (a better approach; still the explanatory note is superfluous; why rule out the Big Mac Index?)

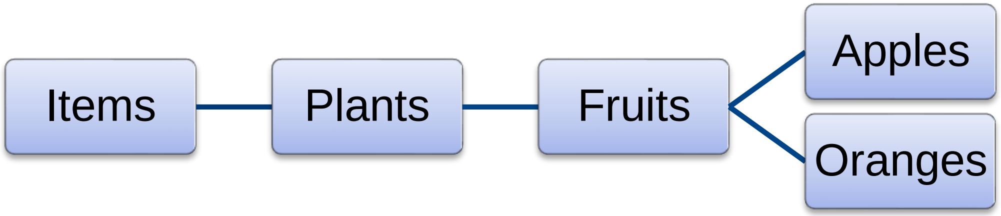
Generic units of measure vs change in units of measure

Indicator	Unit of measure
Debt to GDP ratio	Percent
Debt	National currency
GDP	National currency

Indicator	Unit of measure
Debt	Percent of GDP
Debt	National currency
GDP	National currency

Counting

- Generic vs specific units of measure, e.g. entities vs apples
- A full spectrum of options may exist from generic to specific:



- Leaning towards 'specific', but in practice should be determined by intended uses ('specific' is less parsimonious, more limiting in operations – avoids mixing apples and oranges)

Some recurring modelling choices

- Modelling exactly or modelling functionally (can the two co-exist?)
  - [Age group examples: Youth, Working age vs. Y_LE15, Y15T64: consequences on scope, attributes usage, alignment of data, connectivity]

- Coined terms: representing parsimoniously or redundantly?

Indicator	Unit of Measure
Unemployment	Persons
Unemployment	% of active population

Indicator	Unit of Measure
Unemployment	Persons
Unemployment rate	% (of active population?)

And now the floor is open