10TH EXPERT GROUP MEETING ON

Statistical Data and Metadata eXchange

JANUARY 25–28, 2021

Comparison of Modeling Approaches

January 26, 2021

Abdulla Gozalov
United Nations Statistics Division

sdmx
Statistical Data and Metadata eXchange

Common challenges of modeling data exchange

- Many stakeholders with differing priorities
- Large number of concepts
- Multiple classifications in use
- Sparse data universe
- Uncertainty over the shape of data at the beginning of the project
- Need for custom disaggregation on the part of data reporters
- No universally accepted approach to complex data modeling or criteria for testing a data model
  - No 3NF for data exchange!

Questions designing a data model

- How complete is our knowledge of the data domain?
- One DSD or many?
- Clean or mixed dimensions?
- Should a concept be a dimension or attribute?
- How often are code lists expected to change?

- Guidance for the Design of Data Structure Definitions and Modelling Statistical Domains in SDMX are very useful and offer considerations and possible approaches to the design of DSDs.
- In this presentation, we will compare approaches taken by 3 data exchange initiatives: **Census Hub**, **EcoFin**, and **SDGs**.

1. 2011 European Census Hub

- Developed for exchange and dissemination of EU Member States’ census data
- Hypercubes defined by EU legislation
  - HC01 Total population by geography, sex, household status, legal marital status, country/place of birth, country of citizenship, age
  - HC02 Total population by geography, sex, household status, educational attainment, country/place of birth, country of citizenship, age
  - HC03 Total population by geography, sex, age, household status, status in employment, country/place of birth, country of citizenship, age

    …
  - HC60 Number of all conventional dwellings by geography, occupancy status and type of building
- A total of 60 hypercubes
- 45 variables

Census Hub: Data model

- **“Pure”** approach
- One DSD per hypercube → **60 DSDs**
- Clean dimensions
  - No *Not Applicable* codes
- Dense hypercubes

No.	Breakdowns	DataFlowID
1	GEO.L. SEX. HST.H. LMS. CAS.L. POB.L. COC.L. AGE.M.	HC01
1.1.	GEO.L. SEX. HST.H. LMS. AGE.M.	HC01
1.2.	GEO.L. SEX. HST.H. LMS. CAS.L. POB.L.	HC01
1.3.	GEO.L. SEX. HST.H. LMS. CAS.L. COC.L.	HC01
1.4.	GEO.L. SEX. HST.H. CAS.L. AGE.M.	HC01
1.5.	GEO.L. SEX. HST.H. POB.L. AGE.M.	HC01
1.6.	GEO.L. SEX. HST.H. COC.L. AGE.M.	HC01
2	GEO.L. SEX. HST.H. EDU. CAS.L. POB.L. COC.L. AGE.M.	HC02
2.1.	GEO.L. SEX. HST.H. EDU. AGE.M.	HC02
2.2.	GEO.L. SEX. HST.H. EDU. CAS.L. POB.L.	HC02
2.3.	GEO.L. SEX. HST.H. EDU. CAS.L. COC.L.	HC02
2.4.	GEO.L. SEX. HST.H. CAS.L. AGE.M.	HC02
2.5.	GEO.L. SEX. HST.H. POB.L. AGE.M.	HC02
2.6.	GEO.L. SEX. HST.H. COC.L. AGE.M.	HC02
3	GEO.L. SEX. HST.H. SIE. CAS.L. POB.L. COC.L. AGE.M.	HC03
3.1.	GEO.L. SEX. HST.H. SIE. AGE.M.	HC03
3.2.	GEO.L. SEX. HST.H. SIE. CAS.L POB.L.	HC03
3.3.	GEO.L. SEX. HST.H. SIE. CAS.L. COC.L.	HC03

2. IMF Economic and Financial Statistics

- **EcoFin** – an SDMX data structure for data dissemination and exchange
- Developed by the IMF to support SDMX dissemination of data covered by the IMF "Data Standards Initiatives" (DSIs)
  - DSIs cover a very large array of economic and financial statistics
- A single DSD to support countries data dissemination of macroeconomic and financial statistics using SDMX.

EcoFin Data Model

- **“Simple”** approach
- A single DSD
- 5 Dimensions
  - FREQ
  - REF_AREA
  - INDICATOR
  - COUNTERPART_AREA
  - DATA_DOMAIN
- INDICATOR: mixed dimension
  - Includes all breakdowns except frequency, geography
  - **>65,000 entries in the code list**
- Dense hypercube

CL_INDICATOR	
FCIODC_NUM	Financial, Financial Access Survey, Geographical Outreach, Number of Institutions, Other Depository Corporations, Commercial banks, Number of
FCIODU_NUM	Financial, Financial Access Survey, Geographical Outreach, Number of Institutions, Other Depository Corporations, Credit unions and financial cooperatives, Number of
FCIODMF_NUM	Financial, Financial Access Survey, Geographical Outreach, Number of Institutions, Other Depository Corporations, Deposit taking microfinance institutions (MFIs), Number of
FCIODD_NUM	Financial, Financial Access Survey, Geographical Outreach, Number of Institutions, Other Depository Corporations, Other deposit takers, Number of
FCIOFM_NUM	Financial, Financial Access Survey, Geographical Outreach, Number of Institutions, Other Financial Corporations, Other financial intermediaries, Number of
FCIOFMFN_NUM	Financial, Financial Access Survey, Geographical Outreach, Number of Institutions, Other Financial Corporations, Other financial intermediaries, of which: non-deposit taking microfinance institutions (MFIs), Number of
FCIOFI_NUM	Financial, Financial Access Survey, Geographical Outreach, Number of Institutions, Other Financial Corporations, Insurance corporations, Number of
FCBODC_NUM	Financial, Financial Access Survey, Geographical Outreach, Number of Branches, Excluding Headquarters, Other Depository Corporations, Commercial banks, Number of

Statistical Data and Metadata eXchange

3. Sustainable Development Goals

- UN General Assembly Resolution A/RES/70/1 calls for *"…data which is high-quality, accessible, timely, reliable and disaggregated by income, sex, age, race, ethnicity, migration status, disability and geographic location and other characteristics relevant in national contexts."*
- SDG Indicator Framework approved in support of the Sustainable Development Goals programme in March 2016
- 232 indicators
  - Many more "sub-indicators", or Series
- Uncertainty over disaggregation availability, composition, and frequency of occurrence
- Need to support custom disaggregation in countries

SDG Data Model

- Mixed approach
- A single DSD for all indicators
  - To improve ease-of-use and response rate
- **16 dimensions total**
- **3 mixed dimensions** that include more than one breakdown
  - Series
  - Composite Breakdown
  - Custom Breakdown
- Sparse hypercube
- Guidance for the customization of the DSD for national indicator frameworks available

SDG DSD: Mixed Dimensions

CL_SERIES	
SH_ACS_UNHC	Universal health coverage (UHC) service coverage index [3.8.1]
SH_XPD_EARN25	Proportion of population with large household expenditures on health (greater than 25%) as a share of total household expenditure or income [3.8.2]
SH_XPD_EARN10	Proportion of population with large household expenditures on health (greater than 10%) as a share of total household expenditure or income [3.8.2]
SH_AAP_ASMORT	Age-standardized mortality rate attributed to ambient air pollution [3.9.1]
SH_AAP_MORT	Crude death rate attributed to ambient air pollution [3.9.1]
SH_HAP_ASMORT	Age-standardized mortality rate attributed to household air pollution [3.9.1]
SH_HAP_MORT	Crude death rate attributed to household air pollution [3.9.1]

Cl_COMP_BREAKDOWN	
HZT_WLDFR	Hazard Type: Wild Fire
HZT_WNDST	Hazard Type: Windstorm
MOT_AIR	Mode of Transport: Air
MOT_RAI	Mode of Transport: Rail
MOT_ROA	Mode of Transport: Road
MOT_IWW	Mode of transport: Inland waterway transport
MOT_SEA	Mode of Transport: Maritime
IHR_01	IHR Capacity: National legislation, policy and financing
IHR_02	IHR Capacity: Coordination and National Focal Point communications

CL_CUST_BREAKDOWN	
_T	No breakdown
C01	Custom code 01
C02	Custom code 02
C03	Custom code 03
C04	Custom code 04
C05	Custom code 05
C06	Custom code 06

Data Model Comparison

	Census Hub	EcoFin	SDGs
Type of approach	Pure	Simple	Mixed
Number of DSDs	**60**	1	1
Number of dimensions	5-8 per DSD	5	**16**
Of which mixed dimensions	0	1	**3**
Geography type	Subnational	National	National
Largest code list	>11,000 (geography)	**>65,000 (indicator)**	>600 (geography)
Hypercube(s)	Dense	Dense	Sparse

Conclusions

- Complexity inherent in highly multi-dimensional datasets will manifest itself one way or another
  - Large number of data structures
  - Large code lists
  - Large number of dimensions
  - Complex, mixed dimensions
  - A combination of the above
- There are pros and cons associated with each approach
- Mitigating the complexity is an important consideration

THANK YOU!