Modelling approaches Census Hub

January 26, 2021

Luca Gramaglia
Team Leader – Data and metadata exchange standards
Eurostat – European Commission

sdmx

Statistical Data and Metadata eXchange

Census Hub - Context

- In 2008, the European Commission established common rules for the decennial provision of population Census data (Regulation 763/2008). The first census round occurred in 2011, and the next one will take place in 2021.

- The 2011 and Census 2021 regulations specify a set of "hypercubes" (= multi-dimensional tables containing several breakdowns) to be provided. Within each hypercube, certain subsets were mandatory, while others were optional.

- Large number of hypercubes:
  - 2011: 60 hypercubes
 - 2021: 41 hypercubes

sdmx
Statistical Data and Metadata eXchange

Census Hub – Technical approach
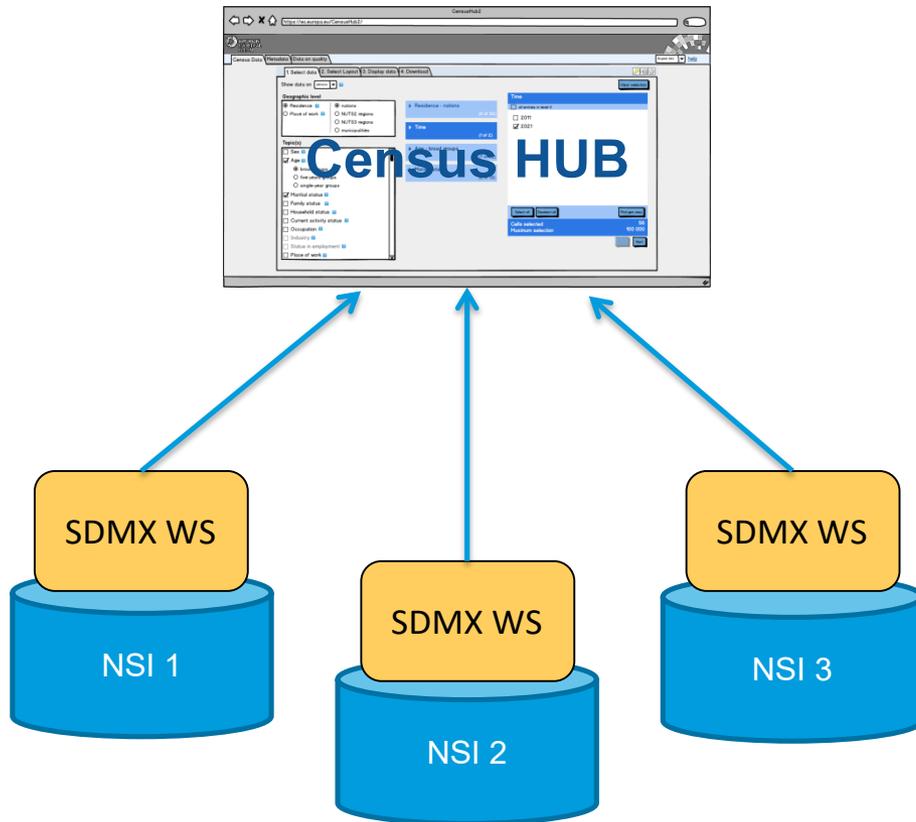
"Hub" approach:

- Each EU Member States exposes hypercubes via SDMX web services. Central Hub pulls data from national web services.

- Avoids very large data transfers from member States to Eurostat.

- Member States retain higher control over data they choose to expose

sdmx
Statistical Data and Metadata eXchange

Census Hub – Modelling approach
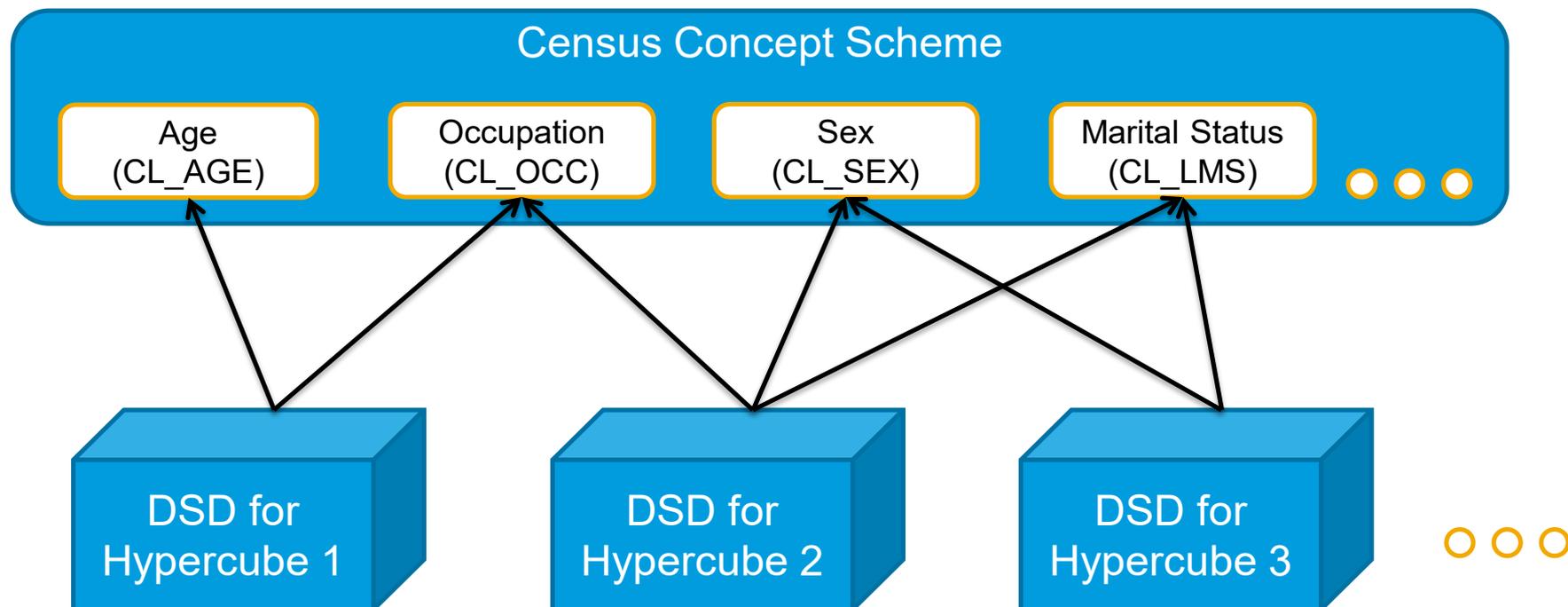
- One common set of concepts and code lists, without "composite" concepts
- A very large number of DSDs (one per hypercube)

Why this approach?

Requirement	Modelling implications
Census regulation clearly specifies data requirements (hypercubes)	A close alignment between the data modelling and the regulation would increase clarity and avoid misunderstandings
Requirements defined in a regulation and updated every 10 years – very low likelihood or needing to extend the hypercubes on short notice	"Tight-fitting" DSD structures that do not allow for easy inclusion of new breakdowns are "good enough"
Hypercubes defined in regulation already expected not to be densely populated (only certain parts are mandatory)	A separate DSD per hypercube avoids having a data model that is very sparsely populated and where users need to "fish" for data.

One DSD per hypercube

sdmx
Statistical Data and Metadata eXchange

Limitations and mitigation actions

Maintenance must be a nightmare – so many DSDs to manage and update!

It would be if we needed to change the structure more than once every 10 years!

What if you want to extend an hypercube with a new breakdown? Your DSDs cannot be easily extended.

Once again, if the structure is set in stone legally then we're pretty safe from that point of view

Your users must hate you – how do they know in which hypercube to look for the data they are interested in?

Well, depending on how you design the Hub you can hide that complexity…

sdmx
Statistical Data and Metadata eXchange

Limitations and mitigation actions

- In Census Hub application users select breakdowns they want – the Hub then selects the appropriate hypercube

- Users do not see the main complexity of the model (high number of DSDs)

Would we do it the same way again?

- Census Hub was designed back in 2009-2010 – back then, SDMX 2.0 was the latest version of the standard.

- SDMX 3.0 will introduce certain features that may change the balance in the equation over which modelling approach is best:

  ➢ On the one hand, features like semantic versioning may drastically lower the maintenance burden related to maintaining a high number of DSDs

  ➢ On the other hand, Constraints are now available in SDMX 2.1 and will be improved SDMX 3.0, and they allow the clear specification of data requirements without resorting to a high number of "tight-fitting" DSDs

¯_(ツ)_/¯

sdmx
Statistical Data and Metadata eXchange

Discussion [Q&A]

For follow-up questions:
luca.gramaglia@ec.europa.eu