Discussion paper: IR on data interoperability – Change proposal(s) on the data models of Population Distribution & Demography (PD)

Detailed issues addressed
All issues related to the data model of Population Distribution & Demography (PD) and how it links to Statistical Units (SU) (section 2.3 in the change proposal summary document). Some of the issues also refer to similar issues in the data model of Human Health and Safety (HH).

Summary of the issue(s)
The data models of PD basically specified a StatisticalDistribution spatial object type as a data cube with a number of dimensions. One of these dimensions is a reference to a statistical unit (through its INSPIRE or thematic identifier), while the others are classifiers (e.g. age group, gender, ...). The only geometry available in the StatisticalDistribution is the areaOfDissemination of the StatisticalDistribution, which represents the overall area of the data set. No common GIS tools can directly use or extract the population distribution data from this data cube representation. On the other hand, it closely mirrors the structure used in the statistical community’s standard format SDMX.

To make population data more easily or directly usable in GIS tools, one of the following two representations might be more adequate.

(1) StatisticalUnit subtype with complex population attribute
In this option, a PopulationDistributionUnit is defined as a sub-type of StatisticalUnit (probably there would need to be one for the Vector and one for the grid representation of SU), which would have 1..* population values, one for each combination of associated classifiers. For example, if no classifier is associated, there would be just one value (the total population for the unit), or there could be one value per age group (or gender), or one value per combination of age group and gender. In this option, the data set would contain spatial objects (the PopulationDistributionUnit) directly usable in a GIS, but the population values would still be provided in a complex data structure, which would not be directly accessible by most GIS systems.
(2) **StatisticalUnit** subtype with simple population attribute

Another solution would be to define a *PopulationDistributionUnit* as a sub-type of *StatisticalUnit*, which would have exactly 1 population value. In this case there would need to be one data set (or layer) of such *PopulationDistributionUnit* objects per combination of relevant classifiers, i.e. one data set for population for all females of ages 0-10, one data set for all males of ages 0-10 etc. The used classifiers should be described in another object (e.g. called *StatisticalDistribution*) that could also serve to group the individual data sets/layers together. In this option, the data set would contain spatial objects (the *PopulationDistributionUnit*) with population values directly usable in a GIS, but there would be a lot of different data sets / layers.

For each of these two options, the population can of course be extracted from the original data cube through slicing (i.e. a query reducing the dimension of the data cube), and then joining it to the geometry of the referenced statistical unit. This is why several issues propose the possibility to use an OGC **Table Join Service (TJS)** to create view and download services (OGC WMS and WFS, respectively). This would allow data providers to make available their PD data sets in representations...
that are more GIS friendly (such as the two options discussed above), even if they are internally represented using a data cube model.

While the publication of such additional representations (e.g. using a TJS) is clearly useful for GIS users, the basic question is whether this is sufficient to meet the current IR requirements or whether the PD data would still need to be published also in a data cube representation following the structure of the data model in the IR.

Proposed change(s)
The proposed changes are not always clear. Most issues refer to the possibility of using a TJS as described above and/or of using SDMX as an alternative encoding.

Possible options
Preferred option (for discussion at MIG-T meeting #48): option 1b

Option 1a: No IR changes, additional representations as a Good Practice
- **Description**: The PD data model is kept as it is in the IRs (except for possibly a few minor bug fix changes). Additional encodings (incl. SDMX) are defined as a Good Practice that can be used to represent the PD in a more GIS-friendly way in addition to publishing them strictly following the current IR data model. This should include guidance on how to use the TJS (and/or offline transformation/ETL processes) for creating such additional representations.
- **Impact**:
  - There will be no impact on the IR or the TG.
  - There will be little impact on existing implementations, but there will be additional effort to publish the PD data in different representations.

Option 1b: No IR changes, additional representations in the Technical Guidance
- **Description**: The PD data model is kept as it is in the IRs (except for possibly a few minor bug fix changes). Alternative encodings (incl. SDMX) are defined in the Technical Guidance, including a mapping from the IR model to these alternative encodings, so that PD data can be published either in a more GIS-friendly way and/or strictly following the current IR data model. This should include guidance on how to use the TJS (and/or offline transformation/ETL processes) for creating such additional representations.
- **Impact**:
  - There will be no impact on the IR, but the TG needs to be updated.
  - There will be little impact on existing implementations, but PD data will be published in different representations (GIS objects or data cubes), which might reduce interoperability. This drawback could be mitigated by defining one default encoding (e.g. SDMX) and recommending other encodings as additional ones.

Option 2: IR amendment modifying the PD data model
- **Description**: The IR will be amended to replace the current data model with a more GIS-friendly one (options 1 or 2 discussed above). The TG, data model repository and schemas will need to be updated accordingly.
- **Impact**:
  - This would be a considerable change in the IR, TG, data model and schemas.
  - There will be considerable impact on existing implementations, which would have to be updated to fit the new model. All new implementations would be more easily usable in GIS systems, and would be interoperable among them.