Integration of INSPIRE & SDMX data infrastructures for the 2021 population and housing census

Nadezhda VLAHOVA, Fabian BACH, Ekkehard PETRI[*], Vlado CETL, Hannes REUTER – European Commission
(*ekkehard.petri@ec.europa.eu)

Abstract: Geospatial and statistical data and metadata are shared using different data formats, exchange methods and dissemination standards. In Europe, geospatial information is shared using the spatial data infrastructure INSPIRE while statistical information is exchanged following Standard for Data and Metadata eXchange (SDMX). Defining a mapping between these two standards is essential to support the combination of these two types of information and thus create greater value for the data and produce integrated datasets. This would allow statistics organisations to enrich their datasets with geospatial information and re-use INSPIRE enabled tools and services to illustrate statistics in a visually enhanced manner. Eurostat carried out a successful pilot study to integrate INSPIRE concepts into SDMX in the context of the Census 2021 data collection. A mapping between the INSPIRE themes Population distribution and Statistical Units on the one hand and Census data and metadata modelled in SDMX for the exchange of census information on the other was defined and will be implemented in Eurostat's SDMX enabled data exchange infrastructure, the Census Hub. As a result Statistical Offices will implement the requirements from INSPIRE without disruption of their established production systems and without double data sharing burden.

1. INTRODUCTION

In the European Union, geospatial and statistical data and metadata are shared using different data formats, exchange methods and standards. Geospatial information is shared using the spatial data infrastructure INSPIRE[1] while statistical information is exchanged using the standards for Data and Metadata eXchange[2] (SDMX).

According to the INSPIRE implementation roadmap, census information needs to be shared in an INSPIRE compliant manner as of 2021. This means that for the 2021 round of population and housing censuses in the EU, statistical offices will have to share census data according to INSPIRE legislation in addition to the existing statistical dissemination infrastructure based on SDMX.

As a result of this double obligation on Member States the goal for the EU wide 2021 census has been to minimise the effects on Member States and to maximise the usability of the census information for the statistical and geospatial community. In this paper we describe the requirements for data sharing of census information from an INSPIRE perspective and present a solution that ensures that Member States meet the legal requirements of INSPIRE while reusing existing tools and standards to a maximum extent.

2. INSPIRE REQUIREMENTS

The INSPIRE directive and its implementing regulations require public authorities in Member States holding spatial data to share these data via national spatial data infrastructures (NSDI). Population distribution is one of the themes that is covered by INSPIRE and as a consequence Statistical Offices holding census information have to meet the legal requirements of INSPIRE. Specifically NSIs need to meet the following requirements:

1) Transformation of population data into an INSPIRE data model (data interoperability);

2) Provision of network services:

    a) Discovery service to search for data and services by means of INSPIRE metadata;

    b) View services to view the data;

---

[1] Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE) http://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:32007L0002

[2] https://sdmx.org/

c) Download service to obtain copies of the data;

3) Creation of INSPIRE metadata on population grid data and the above services;

The INSPIRE roadmap requires Member States to fully comply with INSPIRE by the end of 2021.

3. CENSUS HUB

The Census database is the result of a major joint effort by the European Statistical System (ESS) to better disseminate the results of the Population and Housing Censuses in Europe, providing users with easy access to detailed census data that are structured in the same way and methodologically comparable between countries.

The "Census Hub[3]" tool constructed for data dissemination is based on the concept of data sharing, where National Statistical Institutes (NSIs) provide access to their data according to standard processes, formats and technologies while Eurostat provides the IT structure that allows users to quickly and flexibly specify, compile and extract data stored in the different national census databases. The hub environment has been designed in order to offer an efficient solution for dissemination of census data and metadata. For this purpose, SDMX standards were used. Census data is not previously collected and stored in a central repository but it is directly accessed from the Member States' databases through a central census hub upon request of a data collector (see Figure 1):
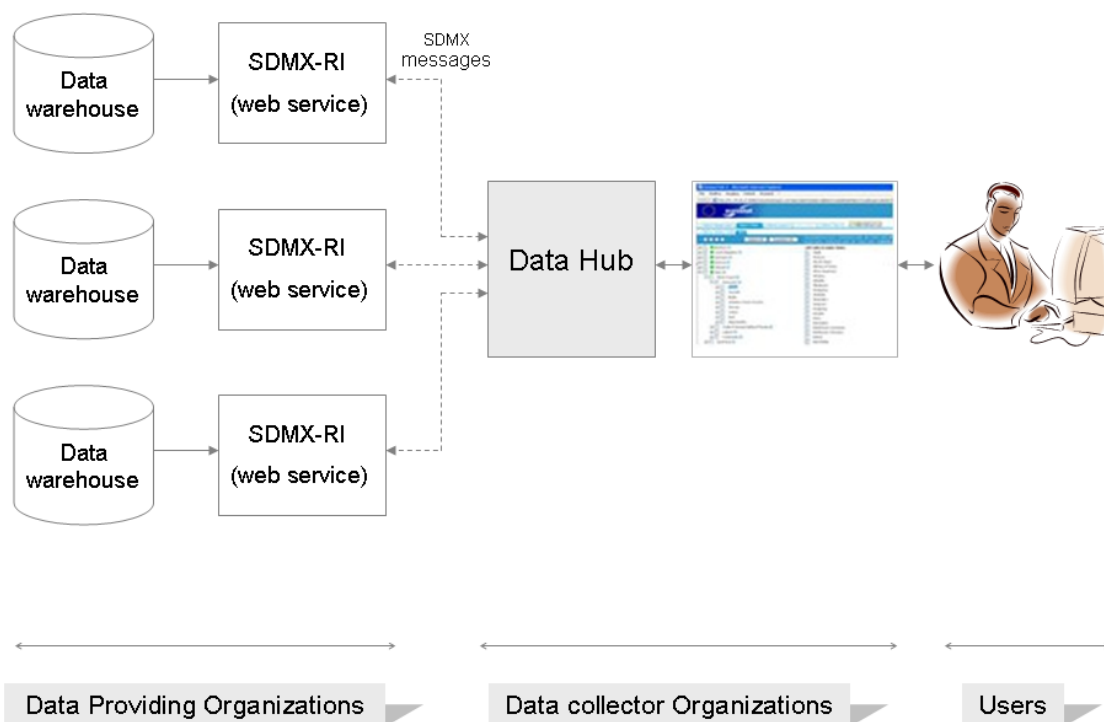

Figure 1: Functionality of the Census Hub.

- A "data user" browses the Hub to define a dataset of interest via structural metadata. Then he chooses the organization of the output layout;

---

[3] http://ec.europa.eu/eurostat/documents/4031688/6285607/KS-02-14-480-EN-N.pdf/05b4ca91-1f72-4dbb-ae2c-d3a07f56d795

- The Hub converts the user request into an SDMX Query and sends the SDMX Query to an interested NSI's (National Statistical Institute) Web Service;

- The NSI data warehouse sends the result to the NSI web service;

- The NSI Web Service converts the result in a SDMX-ML Data message and sends it to the Hub;

- The same steps are repeated if the user has requested data from different member states;

- The Hub puts together all the SDMX-ML data messages proceeding from the interested NSIs and presents the result to the "data user" in the web browser in readable format.

4. PROPOSED SOLUTION FOR INSPIRE COMPLIANCE OF THE 2021 CENSUS

From the start of the discussion around INSPIRE and the 2021 Census data the similarities between spatial data infrastructures and the Census Hub model have been obvious. The INSPIRE directive and its implementing regulations offer a certain degree of freedom as to how these requirements should be implemented technically, provided that data and metadata are properly structured, formats are documented and the services can be called over the internet and support a number of operations.

To gain a better understanding of possible implementations of INSPIRE for population grid data and to check how existing applications could be reused, Eurostat created a Commission internal expert group consisting of DG Environment as responsible DG for the legal aspects of INSPIRE, JRC as the responsible organisation for the technical aspects of INSPIRE, and units in Eurostat in charge of the Census 2021, the Census Hub and Geographical Information.

The goal of this expert group has been to design a data and system architecture that:

- Fulfils the legal requirements of INSPIRE;

- Is data producer and user friendly both from a statistical and geospatial perspective, mainly by keeping together statistical data and INSPIRE resources in one system;

- Avoids duplication of tools and activities, limits burden on Member States, and as a result is cost efficient.

4.1. Overview of the proposed system architecture

It has been obvious from the outset that the Census Hub data infrastructure has a huge potential to meet these requirements, in particular to avoid duplication of infrastructures.

Therefore the above expert group focussed its assessment on the Census Hub. As regards data interoperability the expert group has focussed on the INSPIRE and SDMX implementations of data models for population grid information, thereby for the moment leaving out a more generic integration of INSPIRE and SDMX: The goal of the analysis has been to confirm that a central implementation of the INSPIRE requirements in the Census Hub:

- is legally compliant;

- works technically;

- requires little investments by reusing existing components of the Census Hub architecture already installed in Member States and available at the Commission level;

The conclusion of the analysis was that the Census Hub could be made INSPIRE compliant with fairly limited effort by using it for the transmission of population grid data and metadata from the

Member States to Eurostat (see Figure 2) in the SDMX data model. The actual implementation of INSPIRE would only happen at the central level in Eurostat.

5. FUTURE DATA ARCHITECTURE FOR INSPIRE COMPLIANT DATA SHARING OF POPULATION GRIDS

The following components exist already and are fully under the control of the European Commission (see also Figure 2):

- ESS Metadata Handler (ESS-MH);
- SDMX Reference Infrastructure (SDMX RI);
- Census Hub central application;
- Eurostat View service capability;
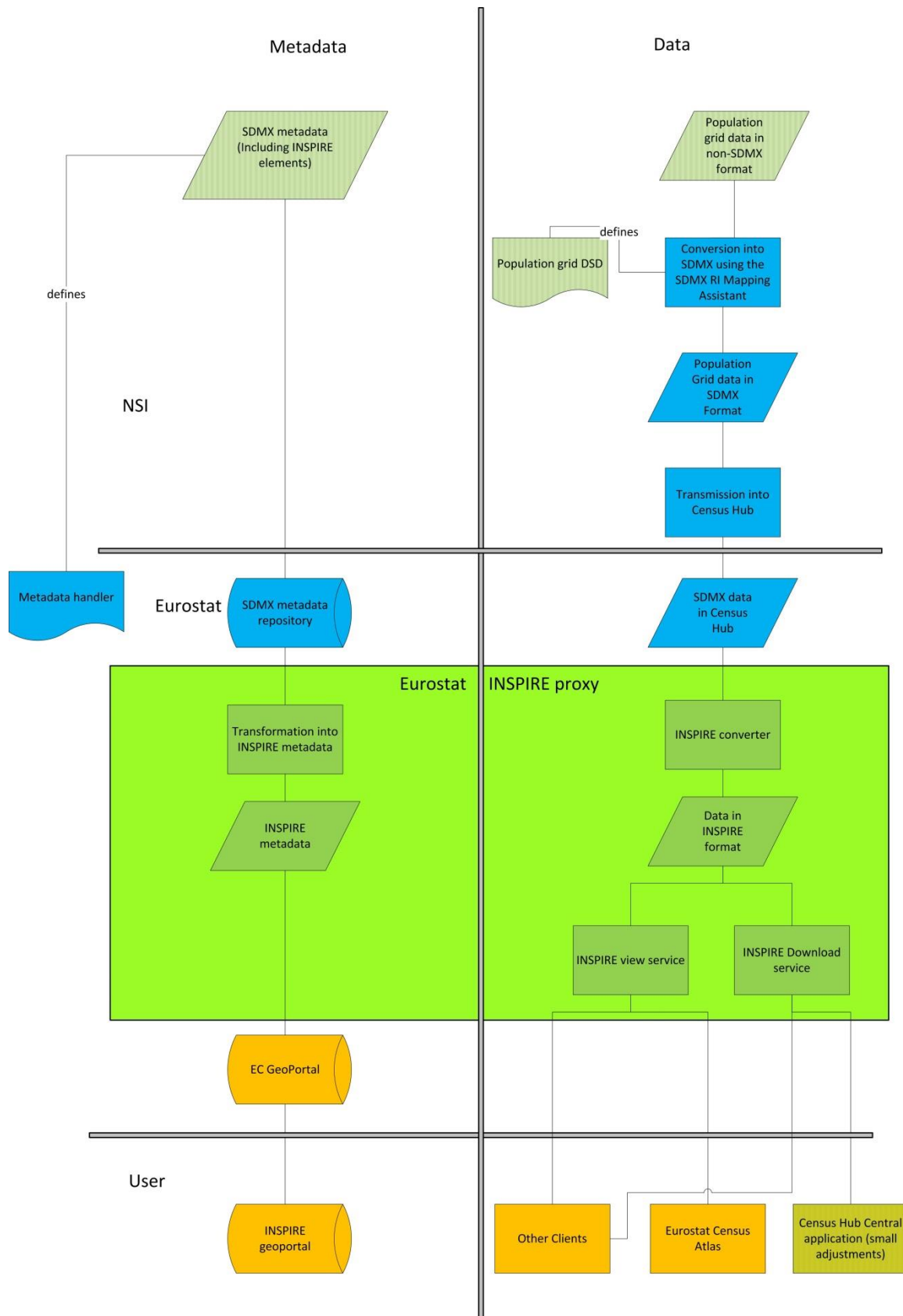- Commission GeoPortal and INSPIRE GeoPortal.

Figure 2: Architecture and workflow for INSPIRE compliant population grids. The orange components are existing INSPIRE components; the blue components are existing Census Hub and ESS Metadata Handler components. The green components need to be developed/ created for INSPIRE compliant population grid data. Hatched components need to be developed for SDMX transfer of population grid data, in parts independently of INSPIRE.

All these components can be reused as is, as they are metadata driven. This means that only a few sub-components need to be added to the current Census Hub architecture, most of which represent in fact regular extensions or use of SDMX:

- Additional reference metadata elements coming from INSPIRE will be added to the ESMS structure in the ESS-Metadata Handler for the Census Hub (see Section 6.1).

- Eurostat will create an SDMX DSD for population grid data. Member States will use the existing data mapper of the SDMX RI to map their original data structure to the Population Grid Data Structure Definition (see Section 6.3).

- Eurostat will develop an INSPIRE extension (INSPIRE proxy) to the Census Hub to offer View Services and Download Services and transform metadata to INSPIRE. This proxy will offer the INSPIRE services to client applications such as the Census Hub central application (for Download) or the Eurostat Census Atlas (for View). In addition the metadata will be incorporated into the INSPIRE geoportal using a Commission INSPIRE discovery service. The presence of INSPIRE metadata in the INSPIRE geoportal is one of the key legal requirements under INSPIRE.

6. MAPPING OF SDMX AND INSPIRE INFORMATION MODELS AND STANDARDS

Eurostat has worked on the mapping of INSPIRE and SDMX concepts for data and metadata and has built Data Structure Definitions (DSD) and Metadata Structure Definitions (MSD) prototypes and examples.

6.1. Mapping INSPIRE metadata to the census ESMS

The implementing regulation of INSPIRE on metadata requires the provision of in total 27 elements to describe the spatial resource (data or service) for discovery in an INSPIRE discovery service. There is general correspondence between *discovery metadata* (INSPIRE) and *reference metadata* (SDMX). Many of the INSPIRE elements are derived from Dublin Core and in essence provide basic information on the title, the history and the content of the resource. As such these elements are very similar to ESMS metadata elements with a similar purpose.

The content that is needed for these 27 INSPIRE metadata elements either had to be covered by existing Census MSD elements or additional elements had to be introduced into this MSD for later extraction by Eurostat for the INSPIRE implementation.

The tasks of matching INSPIRE and SDMX metadata therefore consisted firstly in analysing the semantics of both metadata standards and secondly to establish structural mappings where possible. In a third step, additional elements had to be created for information that could not be mapped into existing MSD elements.

The following cases can be distinguished:

1. The metadata element is the same in terms of semantics and syntax, and a 1:1 mapping can be made;

2. The metadata element cannot be mapped 1:1. This can have several reasons:

   a. The semantics is different, e.g. the scope of an element is narrower or wider in one of the standards;

   b. The encoding of the element is different (e.g. free text vs code lists).

3. The information from INSPIRE is entirely missing in the ESMS and needs to be added.

Out of the 27 metadata elements in INSPIRE, a few mainly on resource identification could be directly mapped to census MSD elements. To avoid complex partial mappings due to undercoverage, overcoverage or different syntax, the remaining elements were simply added as additional INSPIRE concepts to the Census MSD. This approach proved to be very successful and a full coverage of all INSPIRE elements in the new, extended Census MSD could be reached.

It should be noted that the mapping of INSPIRE metadata elements to the Census MSD can be easily extended to other statistical domains, as INSPIRE elements are not topic specific. As a result the extended MSD is not restricted to Census data.

One important aspect is that all additional INSPIRE elements that are not yet covered by already existing, equivalent census MSD elements can be filled automatically using existing information. As a result no extra manual work on the side of NSIs related to INSPIRE metadata is expected.

6.2. ESS Metadata Handler

After the extension of the current Census Hub MSD with INSPIRE metadata elements, the transmission of national reference metadata will be supported by the European Statistical System Metadata Handler (ESS-MH) including those additional INSPIRE elements. The ESS-MH is an IT application that allows users (Eurostat and ESS) to produce and disseminate reference metadata files.

More precisely, the ESS-MH web application is a tool intended for the production, transmission and dissemination of national reference metadata (ESS Metadata providers) and the production and dissemination of reference European metadata (Eurostat). This means that NSIs will be able to transmit INSPIRE compliant metadata to Eurostat via the ESS-MH without the need to use INSPIRE encoding or understand INSPIRE metadata.

It will be Eurostat's responsibility to extract the INSPIRE relevant metadata elements from national reference metadata and consolidate and convert them into an INSPIRE compliant metadata file.

6.3. Mapping INSPIRE data models to the Census DSD

The mapping between the INSPIRE and SDMX data models for census statistics has followed a similar approach as for metadata. From the outset SDMX has been an important input for the design of the INSPIRE population distribution and demography data model, and as a result concepts and code lists are aligned to a certain extent. Nevertheless INSPIRE requires additional attributes and the actual naming of attributes is different due to generic INPSIRE requirements.

While at the conceptual level, SDMX and the INSPIRE annex on population distribution and demography are largely aligned, the actual data structures need to be very different mainly due to the structural rigour of SDMX. This proved to be complex but could be solved successfully and there is now an extended Census DSD that accommodates the requirements of population grids under INSPIRE.

Contrary to the metadata, the data model of INSPIRE for census information is theme specific. A transfer of this specific model to other types of statistics is therefore not possible without further analysis.

6.4. Workflow of data and metadata

The following workflow (see Figure 2) provides a more detailed description of what happens where and who does what, starting from the transmission of population grid data from Member States into the Census Hub until the sharing of INSPIRE data and metadata via INSPIRE services to the final users.

1) Member States create population grid data in their preferred format;

2) Member States, using the Population Grid DSD defined by Eurostat, and the data mapping tool of the SDMX RI, transform the data into SDMX format;

3) Member States, using the ESS Metadata Handler provide reference metadata on Population Grids including the additional INSPIRE elements;

4) The Census Hub pulls the population grid SDMX data into a central repository;

5) A converter as part of the INSPIRE proxy developed by Eurostat converts SDMX metadata into INSPIRE metadata;

6) A converter as part of the INSPIRE proxy developed by Eurostat converts SDMX data into INSPIRE data;

7) Eurostat sets up an INSPIRE download service using a local copy of the INSPIRE data. One of the clients of this download service will be the Census Hub central application;

8) Eurostat sets up an INSPIRE view service using a local copy of the INSPIRE data. One of the clients of this view service could be a Census Atlas developed by Eurostat.

9) INSPIRE metadata on data and services will be included into the INSPIRE geoportal via the European Commission INSPIRE discovery service.

7. BENEFITS OF THE SOLUTION

The above solution has a number of benefits for Eurostat, Member States and users:

Member States:

1) do not need to worry about the implementation of INSPIRE regarding the $1km^2$ grid;

2) require very little additional effort for transmitting population grids statistics and as a result incur no additional costs for implementing INSPIRE (only INSPIRE-related additional metadata);

3) obtain a flexible solution that could be reused also for other census statistics.

Eurostat:

1) offers a cost-efficient solution to Member States for making their population grids INSPIRE compliant;

2) benefits from easy maintenance of the solution due to reuse of existing components, standard APIs and protocols;

3) can illustrate statistics in a visually enhanced manner;

4) obtains a test case for interoperability between SDMX and INSPIRE and integration of statistical and geospatial information with potential for later extension;

5) has a pilot implementation for technical data integration that can be extended to other statistical areas.

The user:

1) receives always up-to-date data also at the European level;

2) benefits from one single entry point to all census 2021 data for all reporting countries, including all INSPIRE web services for the 1km2 grid data.

8. CONCLUSION

Defining a mapping between INSPIRE and SDMX data and metadata standards is at the heart of the proposed solution. For statistical offices the implicit integration of INSPIRE requirements into SDMX standards and the central provision of INSPIRE data and services has proven to be the simpler way compared to national implementations. This allows statistical offices organisations to enrich their aggregated datasets with geospatial information and re-use existing tools and services without disruption of their established production systems and without double data sharing burden.

Statistical and geospatial data integration and the actual implementation of INSPIRE standards will happen at the central level via Eurostat and therefore at the end of the statistical production process. For the user this solution is convenient as all information in both standards is available from one central dissemination point.

Future work will have to investigate how this theme specific solution could be enhanced so that it works for all statistical areas and data providers that are concerned by INSPIRE.