

SDMX GLOBAL CONFERENCE

PARIS 2009

THE CENSUS EUROPEAN HUB PROJECT

(Francesco Rizzo, Eurostat)

1. INTRODUCTION

The aim of this document is to provide an overview of the planned Census Hub and the ongoing Census Hub pilot project in order to summarise the issues encountered or identified until now.

Census taking is a very cost intensive exercise justified by the unparalleled quality of the result. An important aspect of that quality is the flexibility to cross tabulate different variables.

The dissemination of the result of the censuses in the European Union should reflect this advantage to the highest possible extent. The objective should be to provide the user with an easy access to detailed census data that are methodologically comparable between the Member States and structured in the same way.

Of course, the level of detail encounters problems of data confidentiality and, in cases where supporting sample surveys are being used, problems of statistical significance (sampling errors). Although these problems occur in all Member States (confidentiality), respectively in many Member States (significance), they differ strongly concerning the aspects influencing them: the legal and technical frameworks of confidentiality protection varies across countries, the control of the sampling error depends on the specific sampling method and the size of the samples.

The notes above have moved the Census Task Force in April 2007 towards a reflexion on how a maximum harmonisation of the census data dissemination could be combined with the maximum flexibility. Of course it was only the first sketch and the feasibility of the project was considered with much further attention in the following months.

2. ANALYSIS OF THE POSSIBILITIES

The Census European Hub is the proposal of a conceptually new system to achieve the dissemination of the 2011 Census data via the Eurostat website¹.

This task could be achieved using two traditional approaches:

- (1) Member States provide microdata to Eurostat. Eurostat aggregates microdata and stores aggregated data in a central repository that will be used by the dissemination system;
- (2) Member States provide predefined tables to Eurostat, and Eurostat simply publishes those tables on its website.

¹ Legal provisions for the dissemination of the 2011 Census data at EU level are set out in the Regulation (EC) No 763/2008 on population and housing censuses (OJ L 218/14).

Approach (1) maximises flexibility in offering data to final users, but has the following drawbacks:

- Aggregation functions on the central system could be very difficult to implement due to:
 - different confidentiality rules to be applied to microdata from different countries;
 - whether data come from a "full" census (conventional or register-based) or from a sample survey.
- Data maintenance could be very cumbersome because every time a revision is issued, an entire set of microdata needs to be updated or replaced.
- It would be necessary to transmit to Eurostat, and store at Eurostat, very large volumes of confidential (and highly sensitive) data.

Approach (2) greatly simplifies the exercise, by overcoming the above drawbacks, but does not offer enough flexibility to final users, who would have limited possibilities to tailor data to their information needs. In addition, NSIs would have to develop specific tabulation procedures to produce the agreed tables.

An intermediate approach could be to agree on a set of predefined and non confidential hypercubes, to be sent to Eurostat and to be used as the base for the dissemination system. This approach solves the confidentiality and sample data problems, in fact it does not need secure transmission and could offer sufficient query flexibility to final users, but still suffers from the maintenance problem and data transmission could be difficult due to the very high number of cells composing the hypercubes.

3. THE IDEA OF A CENSUS EUROPEAN HUB

An alternative approach to the two described in point 2 is the idea of an "information hub", based on the concept of *data sharing*, where a group of partners agree on providing access to their data according to standard processes, formats and technologies.

The hub is a well-accessible system providing involved actors with the following actions:

- Data providers can:
 - notify the hub of new sets of data and corresponding structural metadata (measures, dimension, code lists, etc.);
 - make data available directly from their systems through a querying system.
- Data users can:
 - browse the hub to define a dataset of interest via the above structural metadata;
 - retrieve the dataset from the NSIs.

From the data management point of view, the hub is also based on agreed hypercubes, but here the hypercubes are not sent to the central system. Instead the following process operates:

- (1) a user defines a dataset through the web interface of the central hub using the structural metadata, and requests it;
- (2) the central hub translates the user request in one or more queries and sends them to the related NSIs' systems;
- (3) NSIs' systems process the query and send the result to the central hub in a standard format;
- (4) the central hub puts together all the results originated by all interested NSIs' systems and presents them in a human readable format

This approach obviously overcomes all the above drawbacks and offers the following additional advantages:

- the process allows for complete decoupling of NSIs' systems from the central hub via standard formats and techniques for the exchange of data, metadata and queries;
- NSIs are free to provide more information than what is contained in the agreed hypercubes without additional effort;
- NSIs could use the same infrastructure developed for the Census Hub project to offer other types of data to the outside world.

The Census Task Force, at its meeting in April 2007, agreed that the hub concept potentially offered the most efficient solution to meeting the requirements for dissemination at European level of the 2011 Census data. It was decided to launch a pilot project to test the hub approach and to allow Member States to get experience with the necessary technologies.

4. SDMX STANDARDS SUPPORTING DATA SHARING

The SDMX standards, besides defining standard formats for data and metadata, also define an architecture for data exchange.

SDMX provides guidelines and tools to support the "pull" mode of data sharing, where the collecting organisation retrieves the data from the providers' web servers. The data may be made available for download in a SDMX-conformant file, or they may be retrieved from a database in response to an SDMX-conformant query. In both cases, the data are made available to any organisation requiring them, in formats which ensure that the data are consistently described by appropriate metadata, whose meaning is common to all parties in the exchange.

This architecture often includes also an SDMX registry that implements the general idea of a metadata registry for use with the SDMX standards. In general an SDMX registry acts as back-office application for all others systems. An application which wants a particular dataset, queries the registry to discover where the data are and how to process data and reference metadata correctly.

It is planned to use this architecture to implement the Census European Hub.

5. THE CENSUS HUB PILOT PROJECT

The pilot project started in January 2008 and finished in October 2008 with the following deadlines:

January 2008: start of the pilot project. Four countries decided to participate (Germany, Ireland, Italy and Portugal);

March 2008: preparation of requirement specification, functional and technical analysis;

April 2008: choice of one data hypercube and related breakdowns to use during the pilot; development of the Data Structure Definition (DSD);

June - September 2008: building of application modules (both Eurostat and NSI side); tests;

October 2008: evaluation report of the pilot; functional and technical analysis for the full 2011 Census Hub.

In the Census WG of September 2008 it was proposed to continue the exercise and NSIs were invited to take part. Malta, Slovenia and Czech Republic have already expressed their willingness. Moreover most NSIs have nominated their IT contacts in order to be informed by Eurostat on the developments of the project.

5.1. Characteristics of the hypercube for the pilot

The pilot hypercube was expected to be very simple one in order to let NSIs to produce it in a short period. The data must comprise the following dimensions:

- Sex (M breakdown)
- Age (M breakdown)
- Current Activity Status (M breakdown)

The data provided in response by the NSIs' web services must be allocated to the reference year 2001. They are not expected to be exact, just simulated to ensure a proof of concept.

The breakdowns of the axes are extracted from the Census team document "Breakdowns of the Topics in the EU legislation on Population and Housing censuses".

In addition to the 3 axes of the hypercube, another dimension is made from the geographic selection. This was chosen as simple as possible: one national level, one regional level, and each subsequent level must feature only two sub-areas.

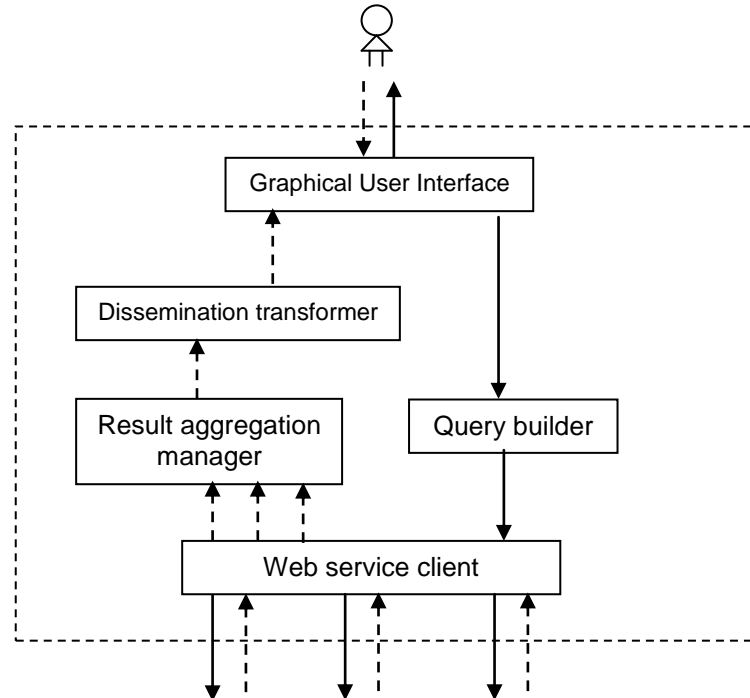
5.2. Architecture

The architecture can be divided in two parts:

- the central Hub, Eurostat side

- the NSI system

The **central Hub** is composed by the following software modules: a *Graphic user interface*, a *Query builder*, a *Web services client*, a *Result aggregation manager* and a *Dissemination transformer*.

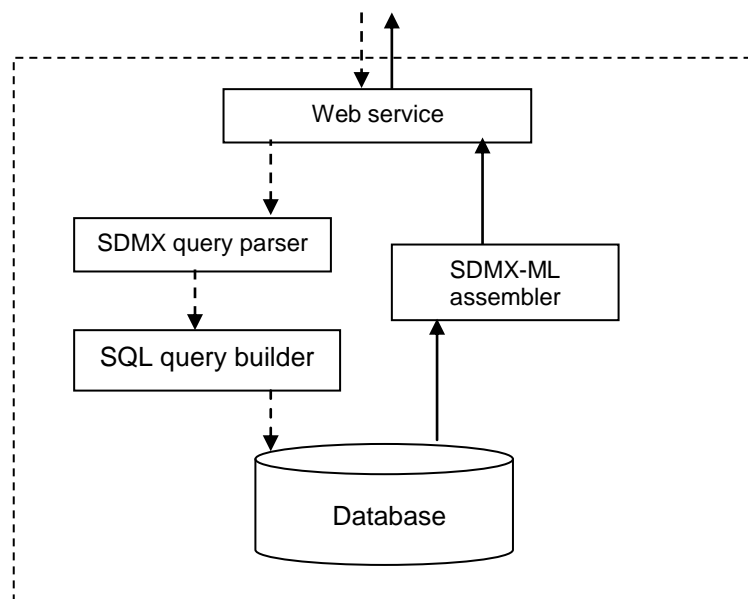


A user starts the flow using the *Graphical user interface*. He/she browses the dimensions and selects a dataset. Then he/she chooses the organization of the output layout specifying which dimension will match X-axis and Y-axis, and which dimension will vary item after item to generate new tables.

Taking into account the user's choices the *Query builder* constructs one or more SDMX queries that will be sent to the related NSIs web services through the *Web service client*.

When the *Web service client* receives the responses (in the format of a SDMX cross-sectional data message) from the queried web services, it forwards those to the *Result aggregation manager*. The *Result aggregation manager* puts together all the received SDMX data messages and sends the result to the *Dissemination transformer* that makes a transformation from an XML format to HTML or CSV.

The **NSI system** is composed by the following software modules: a *web service*, a *SDMX query parser*, a *SQL builder*, a *database* and a *SDMX-ML assembler*.



The *web service* receives a *SDMX* query and forwards it to the *SDMX query parser*. The *SDMX Query parser* breaks down the query and sends it to the *SQL query builder*. The *SQL query builder* creates one or more *SQL* queries and sends them to *Database*. The result is assembled, by the *SDMX-ML assembler*, in a *SDMX* cross-sectional message that will be sent, by the web service, to the central Hub.

6. RESULTS OF THE PILOT

Eurostat has developed the central hub and, at the beginning of February 2009, it will be accessible in a test environment in order to allow all the authorized users in testing the already developed functionalities and the interaction with the peripheral NSIs *SDMX* infrastructures.

Italy, Portugal and Ireland have already setup the architecture while some results from Germany will be expected in the first quarter of 2009.

Italy, Portugal and Ireland have produced documents (available on CIRCA²), regarding their experience during the pilot, which might be used as case studies for the other countries that are preparing in starting the exercise.

Moreover it was produced the *Census Hub Web Service implementation Guidelines*³ that explains how to build web services, using different IT technologies, capable of communicating correctly with the central hub.

² http://circa.europa.eu/Members/irc/dsis/x-dis-xensus-hub/library?l=/census_documents_1/case_studies

³ http://circa.europa.eu/Members/irc/dsis/x-dis-xensus-hub/library?l=/census_documents_1/documents

Finally it is important to highlight how sharing experience and software, between all the involved actors (Eurostat and NSIs), have allowed reducing production costs and development time.

7. BENEFITS IN PARTICIPATING TO THE PROJECT

The whole exercise will let participant NSIs take advantage of Eurostat advice on setting up an SDMX environment and reusing software and experience already developed in other projects. In particular the following benefits will be real:

- participants will be part of a project that will let to share experiences among the different actors, both statisticians and IT personnel, at different levels (planning, production, etc.);
- participants will build an IT infrastructure useful not only for this exercise but also for their 2011 census data warehouse using standards recognized at international level. Moreover the same SDMX architecture, being the most advanced that the SDMX standards describe, could be used in other projects with few or no changes.

8. COSTS IN PARTICIPATING TO THE PROJECT

Costs for implementing an SDMX infrastructure needed for the Census Hub Pilot Project are limited and can be embedded in the more general project that each NSI will support for the 2011 Census.

Particularly the use of an XML-based data format will help to reduce costs of implementation as follows:

- many NSIs are already using, or planning to use XML as the basis for their data management and dissemination systems;
- a wide selection of IT commercial applications and tools are available to work with XML-based data;
- expertise for working with XML is readily available and will often be available in-house;

Moreover, knowledge and software⁴ developed by the participants at the first phase of the pilot are available and can be used immediately.

9. CONCLUSION

The Census Hub pilot project has been necessary in order to well understand how to proceed for the 2011 Census. The used architecture represents the most advanced example of the data sharing detailed in the SDMX standards. Volunteer NSIs can acquire a good experience in managing complex IT projects and a good knowledge of SDMX standards.

Moreover, as the Pilot has been planned as simple as possible in order to let all the NSIs participate with a minor effort, this project is a good occasion for all those who want to start using SDMX.

⁴ http://circa.europa.eu/Members/irc/dsis/x-dis-xensus-hub/library?!=/census_documents_1/available_software/italy