

Distr.  
GENERAL

WP.24  
21 April 2010

ENGLISH ONLY

**UNITED NATIONS ECONOMIC COMMISSION  
FOR EUROPE (UNECE)  
CONFERENCE OF EUROPEAN STATISTICIANS**

**EUROPEAN COMMISSION  
STATISTICAL OFFICE OF THE EUROPEAN  
UNION (EUROSTAT)**

**ORGANISATION FOR ECONOMIC COOPERATION  
AND DEVELOPMENT (OECD)  
STATISTICS DIRECTORATE**

**Meeting on the Management of Statistical Information Systems (MSIS 2010)**  
(Daejeon, Republic of Korea, 26-29 April 2010)

Topic (iv): Innovation and related issues including census systems

## **European Census Hub Project**

Prepared by Adam Wronski, Eurostat and Francesco Rizzo, Istat, Italy

### **I. Introduction**

1. For the first time, the European Union will have a legislation aiming at the availability of harmonised high-quality data from the population and housing censuses conducted in the EU Member States in 2011. The EU legislation on censuses is strictly output oriented: the Member States are free to use the data sources and methodologies they think are best in their countries to produce harmonised census data. They are encouraged to investigate innovative solutions how information from different data sources, including administrative registers, can be linked to provide the required statistics. This way Member States can reduce the burden to respondents and administration, and improve the efficiency of the statistical production of census data.
2. Nevertheless, Census taking remains arguably the most cost intensive exercise in the ESS. It is justified by the unparalleled quality of the results. An important aspect of that quality is the flexibility to cross tabulate different variables, and to provide geographically detailed data.
3. The dissemination of the result of the censuses in the European Union should reflect this advantage to the highest possible extent. The Census Hub project has the main objective of disseminating the result of the censuses in the EU, providing the users with an easy access to detailed census data that are methodologically comparable among the Member States and structured in the same way.
4. Of course, the level of detail creates 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.

5. It should be noted here that the census data to be disseminated are not microdata; they are aggregated data structured according to the hypercubes agreed with Member States and set out in the legal implementing rules for the census Regulation.

6. This situation calls for an innovative technical solution for the transmission and dissemination of census data on the European level. The added value of the census, namely the high geographical resolution and the possibility to cross tabulate harmonised census data, should be offered to the user to the maximum possible extent. Moreover, the volume of census data is particularly high.

## **II. An innovative transmission and dissemination of census data in the EU**

7. The architecture of the Census Hub was presented and discussed for the first time during the Census Task Force in April 2007: after having analyzed different alternatives and the related advantages and disadvantages, the Census Task Force agreed that the Hub approach could offer the most efficient solution to meeting the requirements and the constraints for dissemination at Eurostat level of the 2011 Census data. Moreover 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.

### **A. Analysis of the possibility**

8. The European Census Hub is the proposal of a conceptually new system to achieve the dissemination of the 2011 Census data via the Eurostat website. This task could have been achieved using two traditional approaches:

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

9. To keep the protection of confidential data simple (and this is different from surveys that are not exhaustive) approach (a) was not chosen at this time. Aggregation functions and managing confidentiality would increase significantly the complexity of the project, also due to variability of confidentiality rules among countries.

10. Approach (b) greatly simplifies the exercise but does not offer enough flexibility to final users, who would have limited possibilities to tailor data to their information needs.

### **B. The idea of a European Census Hub**

11. An alternative approach to the two described above 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.

12. A modern dissemination of census data should meet some technical requirements: the dissemination system shall provide the user high data accessibility. Based on harmonised concepts, definitions and specifications of the data transmitted, the tool should be designed to allow maximum flexibility to cross tabulate data. Systematic and standardised metadata should facilitate the interpretation of the data.

13. Despite these tall requests, the dissemination tool should be easy to use. Problems linked to a speedy access to massive amounts of data should be overcome.

14. At the same time, the NSIs appreciate if they remain 'proprietors' of their data and keep complete control over them (data at the source dissemination). The NSIs want to keep the IT platform that they already use for their national purposes. In the case of revisions or updates, it is easier for the NSIs to upload the new data in their own system (something they do anyway) instead of sending a complete new data set to Eurostat.

15. The hub approach offers a very efficient solution to meeting the requirements for dissemination of the 2011 Census data at EU level. The hub is a well-accessible system providing involved actors with the following features:

- (a) 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.
- (b) Data users can:
  - browse the hub to define a dataset of interest via the above structural metadata;
  - retrieve the dataset from the ensemble of NSIs.

16. From the data management point of view, the hub is also based on agreed hypercubes agreed upon in the Draft Regulation on the EU programme of 2011 census data, but here the hypercubes are not sent to the central system. Instead to satisfy user query requirements the following process takes place:

- (1) a user defines a dataset through the web interface of the central hub and 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

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

- (1) 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;
- (2) NSIs are free to provide more information than what is contained in the agreed hypercubes without additional effort.
- (3) NSIs could use the same infrastructure developed for the Census Hub project to offer other types of data to the outside world.

### **III. The Census Hub and SDMX**

18. The SDMX standards, besides defining standard formats for data and metadata, allow defining a particular service infrastructure for data exchange. Each organisation can develop such service infrastructure or its components itself or use all or individual components from other particular service solutions.

19. SDMX provides guidelines and tools to support the "pull" mode of data sharing, where the collecting organisation retrieves the data from the providers' websites. 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. In this case as well one can develop own tools or integrate ready made software is considered advantageous.

20. This service infrastructure 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.

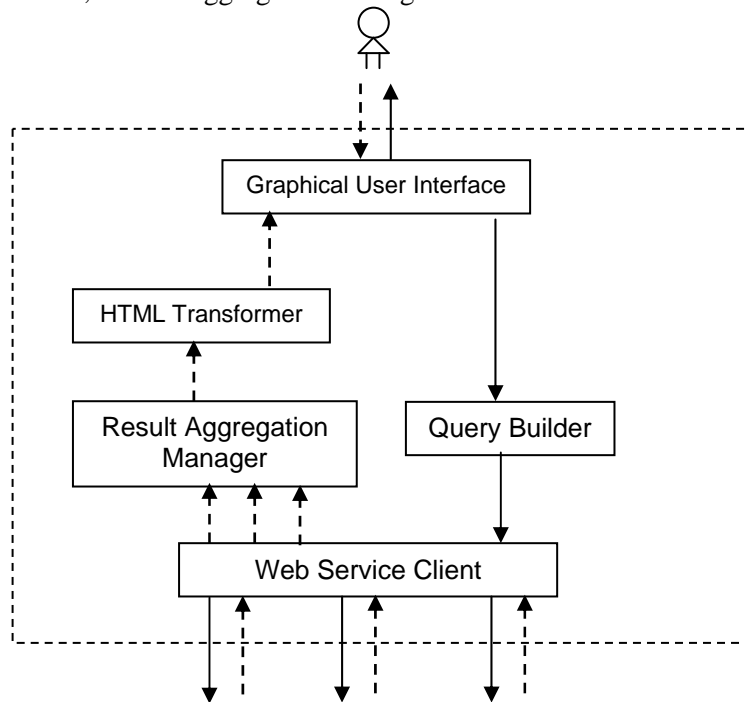
21. It is planned to use this service infrastructure to implement the European Census Hub.

#### A. Census Hub Architecture

22. The full European Census Hub architecture can be divided in two parts:

- the central Hub, Eurostat side
- the NSI system

23. 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 an HTML transformer.

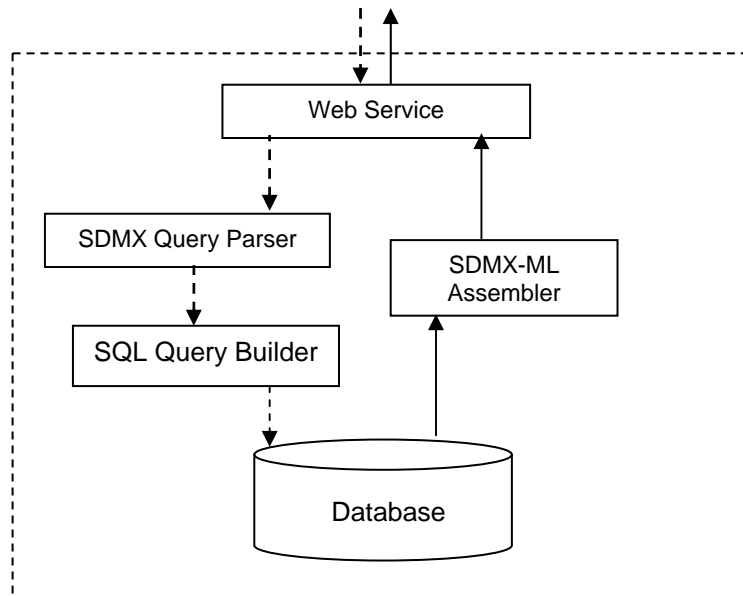


24. A user starts the flow using the Graphical User Interface. He/she browses the dimensions and selects a dataset. Then he/she chooses the organization of the output layout specifying which dimension will match X-axis and Y-axis, and which dimension will vary item after item to generate new tables.

25. 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.

26. When the Web Service Client receives the responses (in the format of a SDMX cross-sectional data streams) from the queried web services, it forwards those to the Result Aggregation Manager. The Result Aggregation Manager puts together all the received data streams and sends the result to the HTML Transformer that transforms from an XML format to HTML and/or other formats.

27. 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.



28. 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 stream that will be sent, by the Web Service, to the central Hub.

#### IV. The Census Hub pilot project

29. The scope of this pilot project is to build the IT architecture that will be hosted in Eurostat (Hub) and facilitate the development of a SDMX IT infrastructure in the Member States. The project aims at three main objectives:

- to develop the Hub Web application that will acts as a client towards MSs' Web services;
- to support the implementation of the MSs' SDMX IT infrastructures through technical advice;
- to facilitate the sharing of software between all the countries involved in the exercise.

##### A. Phase I

30. The first phase of the pilot project started in January 2008 and finished in October 2008. The following dates represent the main deadlines foreseen:

- January 2008: selection of the volunteer NSIs. DE, IE, PT and IT decide to participate;
- March 2008: Requirement specification, functional and technical analysis; choice of one data hypercube and related breakdowns to use during the pilot; development of the Data Structure Definition (DSD);
- June-September 2008: build 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.

##### B. Result of the Pilot Phase I

31. All the involved parties have developed their SDMX infrastructure: Eurostat has developed a prototype of the central hub with the main objectives of testing the “data flow” with the peripheral web services and an appropriate graphical user interface, capable of providing the users with an easy means for accessing data. NSIs have developed their web services capable to allow the access to their data warehouse by external applications, such as the central hub.

32. Each involved Member State has produced a document specifying their experience during the pilot, the support costs and the gained benefits. Those documents are available on CIRCA and could be used by other NSIs as case studies.

33. Eurostat has produced the Census Hub Web Service Implementing Guidelines version 1.0. This document explains how to build web services as part of an overall SDMX service infrastructure, dealing with topics such as which approach to follow when different IT technologies (JAVA and .NET) are used and how to handle errors.

### **C. Phase II**

34. In the Census Working Group of September 2008 the results of the pilot phase I were presented.

35. The main milestones for the pilot phase 2 were the following:

- (a) Involve more Member State in the project. In order to facilitate this process Eurostat launched an action to support SDMX implementation in Member States. The purpose of this action was to provide support to Member States in the area of SDMX, with particular attention to the Census Hub. Currently, it was focused into two directions:
  - providing technical advice in implementing a SDMX IT service infrastructure. At this purpose several technical bilateral meeting had had with interested NSIs (IT, PT, LV, BG, HU, MT, SI);
  - contributing with some open source components to a generic reference SDMX service infrastructure for NSIs.
- (b) contributing with some open source components to a generic reference SDMX service infrastructure for NSIs.
- (c) Develop and test additional functionalities for the central hub

### **D. Result of the Pilot Phase II**

36. Malta, Slovenia, Czech Republic, Bulgaria, Estonia, Spain, Poland, Luxembourg decided to join the pilot project.

37. All the NSIs (with the exception of one country) have nominated their IT contacts in order to be informed by Eurostat on the developments of the project. At this purpose a periodically newsletter has being produced and sent to all NSIs IT contacts.

38. A new version of the Hub application was developed with the following added characteristics:

- New graphical user interface. The AJAX technologies was used in order to optimize the browsing functionalities;
- Cache system. The results of the most used query are stored in a cache to reduce the response time;
- Multi-threading. The web services are accessed in parallel to reduce the response time;
- Filters. The data user has now the possibility to filter the hypercubes of interest using the dimensions as filtering parameters;
- Offline processing. The application is able to manage bulk queries: run in asynchronous mode and the data user can receive the response directly into his web browser or by email.

39. Eurostat developed the SDMX NSI service infrastructure and the related software building blocks as open source packages. Each NSI can re-use all the building blocks as a whole or individually as components to be integrated in an existing dissemination system whenever considered advantageous.

40. Eurostat organized an SDMX Technical Workshop in Madrid attended by 40 IT designers and programmers coming from more than 20 NSIs.

### **C. Phase III**

41. The phase III will consist principally of an enhancement of the overall IT Census Hub architecture, an increase of technical advice to the NSIs, a consolidation of the SDMX service infrastructure developed by Eurostat for the NSIs and a series of capacity building actions that will foster the development of the necessary SDMX service infrastructure in the NSIs. Below the deliverables that will be expected.

42. An alternative Census Hub architecture that will host all those NSIs that will not create their SDMX service infrastructure. In this case data will be fetched in pull mode by the Eurostat Pull Requestor<sup>1</sup> and loaded in a local database. The central Hub will access this local database through a local web service, in the same manner as it access the remote web services in the NSIs' premises.

43. An enhancement of the central application (Hub) with the introduction of the support for the multiple language management.

44. The introduction of a web application that allows the NSIs to manage and update their geographical code list till the LAU 2 level.

45. The launch of a "Census Hub IT Working Group" with the main aim of supporting and coordinating the IT activities.

46. An SDMX technical workshop for IT designers and developers, similar to that held in 2009

### **V. Reusability of the solution, quality and cost benefits**

47. Costs for implementing an SDMX infrastructure needed for the Census Hub project are very limited. 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

48. Moreover pilot phases have been clearly demonstrating that sharing experiences between all the involved actors, both Eurostat and NSIs, and reusing the software developed in other SDMX projects or available in the "SDMX community" could reduce dramatically development costs.

49. At this purpose Eurostat has been actively contributing as follows:

- developed and made available for free-download from the SDMX website an SDMX Infrastructure for NSIs. The source code for this Infrastructure is available so that it can be used as a whole, or as components to be easily integrated into own IT systems in statistical organisations;

---

<sup>1</sup> The Pull Requestor is a module of the Eurostat SDMX Data Repository infrastructure that allows to fetch SDMX data files directly from authorised by NSIs web sites

- financed the ESSNet on SDMX, where some MSs are working to produce software and best practise to be shared with the other countries;
- providing SDMX training organized in two sessions per year, for both statisticians and IT staff.

50. While costs were small, substantial benefits were realised. The following benefits are real and demonstrable:

- The SDMX Infrastructure built for the Census Hub project could be used in other statistical domains with few or no changes.
- If felt beneficial the actual European Census Hub infrastructure could be used for dissemination of other NSI data with the added advantage of using standards recognized at international level.
- Participants are part of a project that will allow sharing experiences among the different actors, both statisticians and IT personnel, at different levels (planning, production, etc.).

## **VI. Conclusion**

51. The Census Hub Pilot Phase 1 has demonstrated the feasibility of the proposed architecture and it was necessary in order to well understand how to proceed for the 2011 Census.

52. The Census Hub Pilot Phase 2 will represent a consolidation of the entire project, both from a technical and a participation point of view.

53. The used architecture represents one of the most advanced examples of the data sharing architecture based on the SDMX standards: volunteer NSIs can acquire a good experience in managing complex IT projects and a good knowledge of SDMX standards.

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

55. As of April 2010, fourteen countries are participating in the pilot project. Three other countries have expressed their interest in joining the project but have not yet made a final decision. Six NSIs have already put in place the SDMX infrastructures.

---