Distr.
GENERAL

Working Paper
11 February 2014

ENGLISH ONLY

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

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

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

**UNITED NATIONS**
**ECONOMIC AND SOCIAL COMMISSION**
**FOR ASIA AND THE PACIFIC (ESCAP)**

**Meeting on the Management of Statistical Information Systems (MSIS 2014)**
(Dublin, Ireland and Manila, Philippines 14-16 April 2014)

Topic (ii): Standards-based modernisation

# European Census Hub: A Cooperation Model for Dissemination of EU Statistics

Prepared by Ioannis Xirouchakis, Eurostat

## I. Introduction

1. It is with Regulation (EC) No 763/2008 of the European Parliament and of the Council of 9 July 2008 and its implementing Commission Regulations that the greatest dissemination effort ever produced by the European Statistical System (ESS) started. This is probably the first time that the ESS makes available so much data at such a level of detail. The 2011 European Census is in line with the United Nations Economic and Social Council's resolution 2005/13, urging countries to carry out a population and housing census at least once between 2005 and 2014 and to disseminate the census results in a timely manner.

2. The aim of the 2011 Census is to provide the public with access to detailed and comparable Census results for all EU Member States[1] and EFTA countries.[2] This will be the first time that Census data for 32 countries are disseminated through a single web site, with the data themselves stored in the dissemination databases of the countries responsible for their production.

3. What makes this Census dissemination special from a technological point of view is the underlying usage of the so-called data Hub. Census data for each country will not be transferred to a central location, but will be exchanged ad hoc between the countries and the Census Hub as soon as end-user queries arrive. Moreover, the Hub will collect queries concerning more than one country, retrieve relevant data and offer the user combined results transparently.

4. In 2007, the Census Task Force, consisting of EU Member States and Eurostat representatives, agreed that the Hub approach could offer the most efficient solution to requirements for dissemination of the 2011

[1] Austria, Belgium, Bulgaria, Croatia, Cyprus, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Ireland, Italy, Latvia, Lithuania, Luxembourg, Malta, Netherlands, Poland, Portugal, Romania, Slovakia, Slovenia, Spain, Sweden, United Kingdom.
[2] Iceland, Liechtenstein, Norway, Switzerland.

Census results at EU level. In 2014, the Census Hub brings a new dimension to cooperation and sharing of data and services at EU level. With the official launch of the interface on Eurostat's website, EU Member States and EFTA countries are making a commitment to ensuring high availability of data and servers, and to providing statistical data of high quality that will be kept available online until at least 2025.

# II. The Census Hub

## A. Rationale

5. For the 2011 Census, countries report 60 data hypercubes, 35 hypercubes with data on quality ('quantitative' metadata), and textual (or 'qualitative') metadata. One of the key elements of the Census regulation is to improve the extent to which the end-user can cross-tabulate various topics (e.g. age, sex, marital status), while preserving a considerable level of detail (also in geographical terms). Analysis has illustrated that the level of detail required by the regulations leads often to hypercubes of eight or more dimensions, with several billion records. The Hub architecture was agreed as the best approach to avoid voluminous data transmissions, as well as to avoid maintaining a central infrastructure of considerable size.

6. The Hub approach allows countries to own and maintain their data, which countries themselves much appreciate. Implementation enables countries to preserve the structure of their production/dissemination databases, as their output is dynamically mapped to an agreed common structure.

## B. State of play

7. The successful development of the Census Hub has been the result of long, persistent efforts. Since 2007, EU Member States and Eurostat representatives have closely collaborated in analysing the endeavour and finally implementing Regulation (EC) No 763/2008 and its three implementing regulations. The Census Hub system and its components have gradually evolved through the project's phases, in order to enable countries to disseminate high quality statistical data flexibly and efficiently.

8. At present, the 2011 Census endeavour is reaching its conclusion. The Census Hub application and tools have been developed, while countries have made their data accessible to the production environment (the so-called Production URL at the European Commission's premises).

9. By the time the Census Hub officially opens in summer 2014, countries will have had the opportunity to confirm that their data are depicted properly and any technical details will have been dealt with. The European Commission (and Eurostat in particular) has committed itself to providing the best possible service and data dissemination that will be both comprehensive and of high statistical quality. In this context, future short-term enhancements to the Census Hub have been planned.

# III. The System

## A. Standardisation

10. An important aspect of the Census Hub's success is standardisation. The standard for Statistical Data and Metadata eXchange (SDMX) has been used, as the de facto standard for the exchange of statistical data and metadata, sponsored by a number of international organisations, including the BIS, the ECB, Eurostat, the IMF, the OECD, the UN and the World Bank.

11. Countries report 60 data hypercubes and 35 hypercubes with data on quality (besides metadata in textual format). In practice, early in the project, these hypercubes and their associated structural metadata (such as dimensions and lists of possible codes per dimension) were described in SDMX terms as 'structure files' (containing SDMX artefacts, such as Data Structure Definitions (DSDs), codelists and geographical

codelists, dataflows, concept and category schemes). The structure files impose an explicit standardisation, consistent throughout the project.

12. Once the structure files are available, experts at country level map the structure of their production/dissemination databases to the DSDs, as well as the codes in use in them to the standard codelists. This mapping, which can in certain cases be complex, is carried out in a straightforward manner using the Mapping Assistant, a tool developed by Eurostat for such purposes. Mapping information is kept in the Mapping Store.

13. Structure files are identical for all countries, with the exception of the geographical codelist, which at country level contains only the codes relevant to that country. Structure files including the full geographical codelist are loaded onto the central node of the Census Hub using the SMD tool, an administration tool developed by Eurostat.

14. As illustrated in the following, at runtime, after the structure files have been loaded on the central node (SMD tool) and on the country nodes (Mapping Assistant) and the mapping is complete:

(a) the central node and the country nodes exchange information (queries and responses) in SDMX format (in fact SDMX-ML messages in cross-sectional data format according to standard SDMX version 2.0) respecting the structure files;
(b) each country node uses the mapping information kept in the Mapping Store to translate SDMX-ML queries into queries that the particular country dissemination database 'understands' (and the reverse for the response).

15. The standardisation approach, together with the development methodology of the Census Hub application, permits structural changes with limited effort. Virtually any structural change is possible after loading the updated structure files on the central node and on the country nodes and executing the mapping to the country's dissemination database. The Graphical User Interface (GUI) of the Census Hub application is then automatically adjusted to reflect the change.

## B. Architecture

16. In short, the Census Hub comprises:

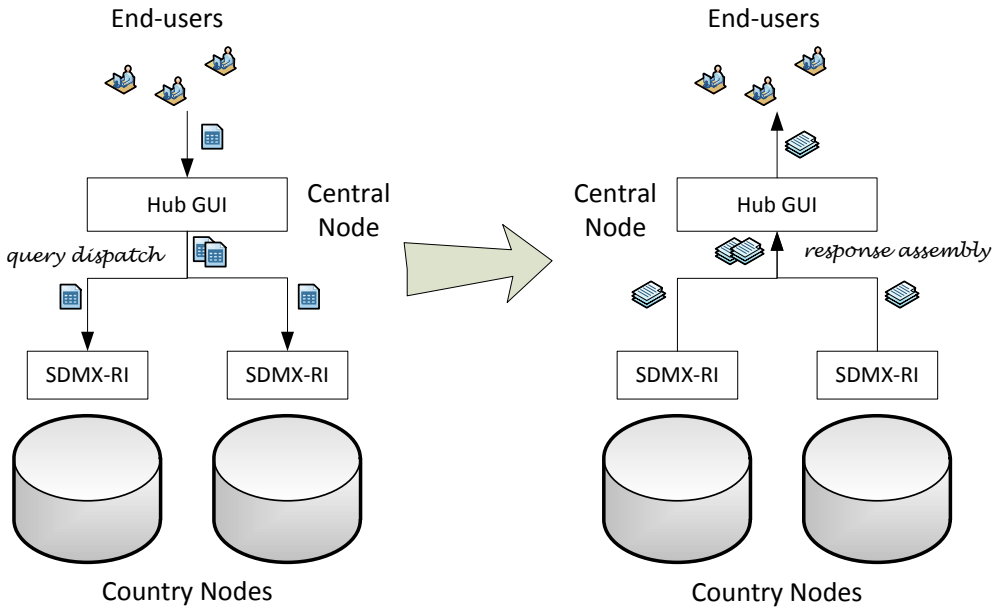
(a) the central node, which hosts the Census Hub application and its various tools and utilities;
(b) the country nodes, which comprise the country's dissemination database and a data mapping and exchange module.

17. Almost all countries participating in the Census Hub project use the SDMX Reference Infrastructure (or SDMX-RI) as a data mapping and exchange module. In practice, the SDMX-RI is a generalised service infrastructure (a set of building blocks) developed by Eurostat that can be re-used partially or completely by any organisation interested in sharing data in SDMX format (for dissemination or data exchange). The SDMX-RI, among other things, parses and validates incoming SDMX-ML messages, translates them into SQL statements by processing mapping information from the Mapping Store and generates SDMX-ML datasets by translating the response of the dissemination databases.

18. In the context of the Census Hub application, when an end-user query is submitted onto the Graphical User Interface (GUI):

(a) the central node formulates it into the appropriate format and dispatches it to the country nodes concerned by the query;
(b) on the side of a concerned country node, the SDMX-RI module transforms the message into a query that the country's dissemination database 'understands';
(c) the country's dissemination database provides a result set in response to the query;

(d) the SDMX-RI module transforms the result set into the SDMX-ML message that the Hub expects to receive and dispatches it;
(e) the Hub collects the result sets, composes and displays them according to the end-user's instructions.




19. The Hub architecture illustrates a number of advantages over traditional approaches:

(a) the end-user is always sure to obtain the latest data revision available;
(b) no voluminous data transmissions to a central location are necessary;
(c) there is no need to store and maintain a large central database;
(d) changes are reflected on the Hub immediately after they are submitted at country level;
(e) countries own their data and assume responsibility for them;
(f) a country can revise its data without impact on the Hub or other country nodes;
(g) countries just need to map their dissemination databases to the provided structure files rather than changing the structure of their databases.

20. The architecture is also advantageous in terms of extensibility:

(a) very limited technical effort is required to add another country to the Hub;
(b) limited effort is required to expand the system to include new datasets (corresponding to new DSDs and codelists);
(c) the same dissemination databases, containing census or other data, can be made available to other Hubs with limited effort (for data mapping mainly).

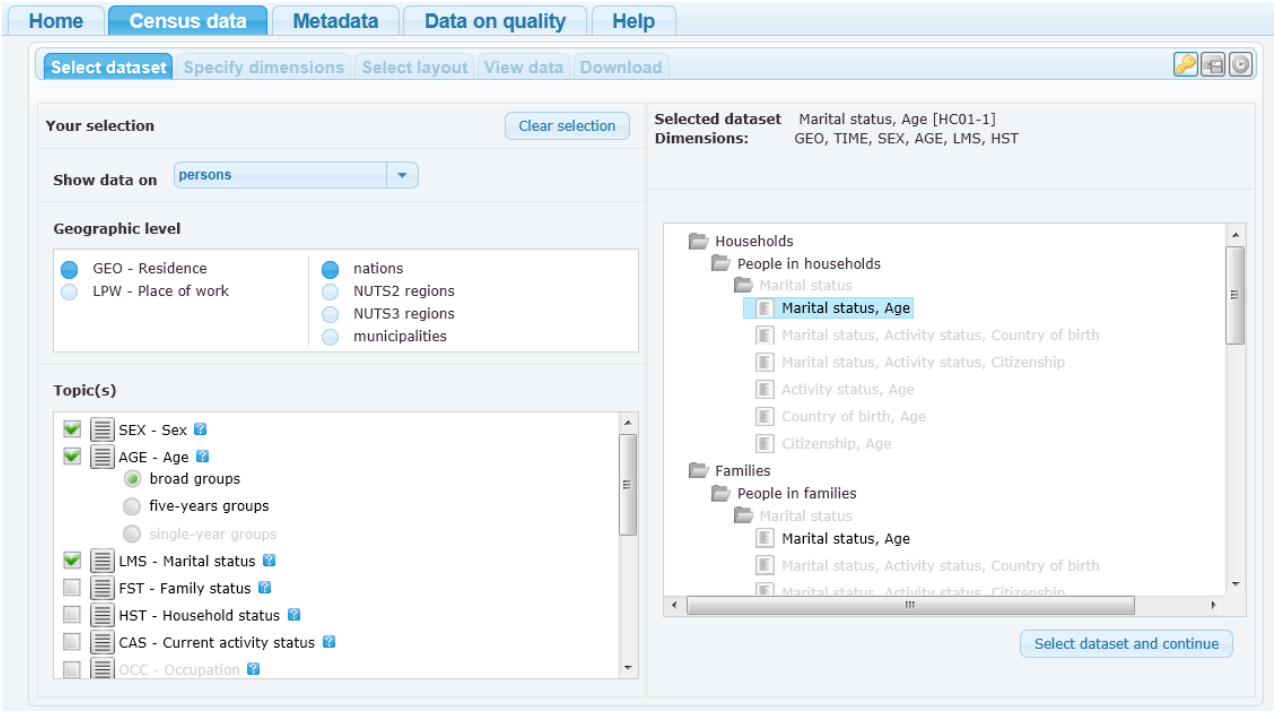
21. The Hub architecture comes with a number of challenges that must be taken into account:

(a) data validation takes place only at country level;
(b) data availability on the Hub depends on country node availability;
(c) response to queries is somewhat slower compared to the traditional approach, as the Hub needs to query and collect responses from country nodes.

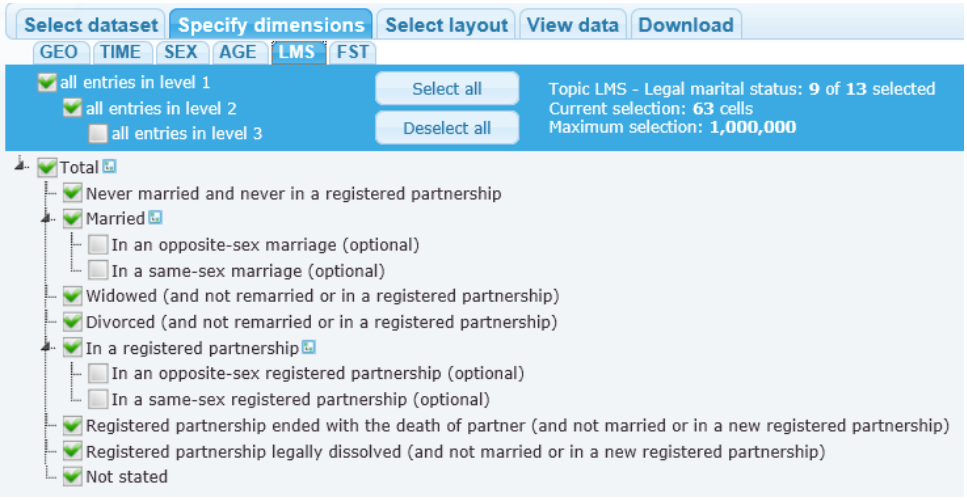
## C. Graphical User Interface

22. The GUI is divided into the publicly available part and the administration tools. The first is meant for the end-user, the second for the administrator and the content manager. The publicly available part

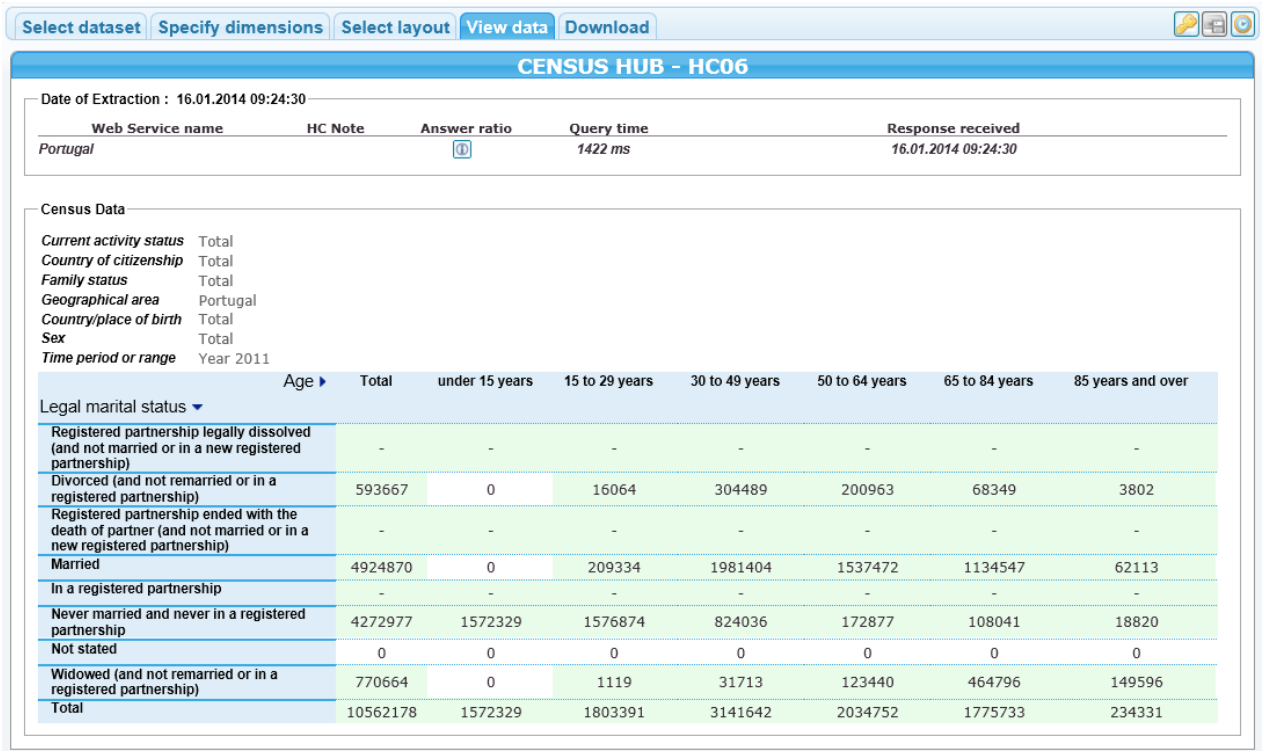
consists mainly of the *Census Data*, *Metadata* and *Data on quality* sections. Under *Census Data*, on the *Select dataset* screen, the user is prompted to select one of the available 60 hypercubes (or 'datasets') on the basis of 'topics' that he/she wishes to cross-tabulate.




23. On the *Specify dimensions* screen, the user is prompted to 'constrain' the hypercubes' dimensions to the desired codes only. Notably, on the geographical (GEO) dimension, the user can go from national (NUTS0) level down to NUTS3 and even municipality (LAU2) level, depending on the hypercube. It is interesting to observe that the screen tabs corresponding to dimensions (e.g. SEX, AGE, LMS) have been created and populated automatically on the basis of the structure files available in the system.




24. On the *Select layout* screen, the user chooses how to visualise the result. The supported currently capability is setting one dimension to the columns and one dimension to the rows of a table; while this table structure is repeated for all possible combinations of the codes in the other dimensions in the *View data* screen. Data appearing in the screenshots hereinafter are test data.




25. Screen *Download* provides the user with the possibility to download the result set in various formats.

26. In a similar manner, under *Data on quality*, the user can query the 35 quality hypercubes, containing numeric assessments of different dimensions related to data quality, such as accuracy, completeness, etc.

27. Under *Metadata*, the user can browse through the textual metadata provided by the countries. These metadata explain the methodology used to produce census data and provide other generic information complementing the data themselves.

28. The administration tools provide the administrator and the content manager of the system with a number of utilities, such us:

(a) configuration of the connections to country nodes;
(b) management of structure files and geographical codelists in particular;
(c) configuration of basic system parameters (e.g. cache clean-up interval);
(d) visualisation of SDMX queries and responses for testing.

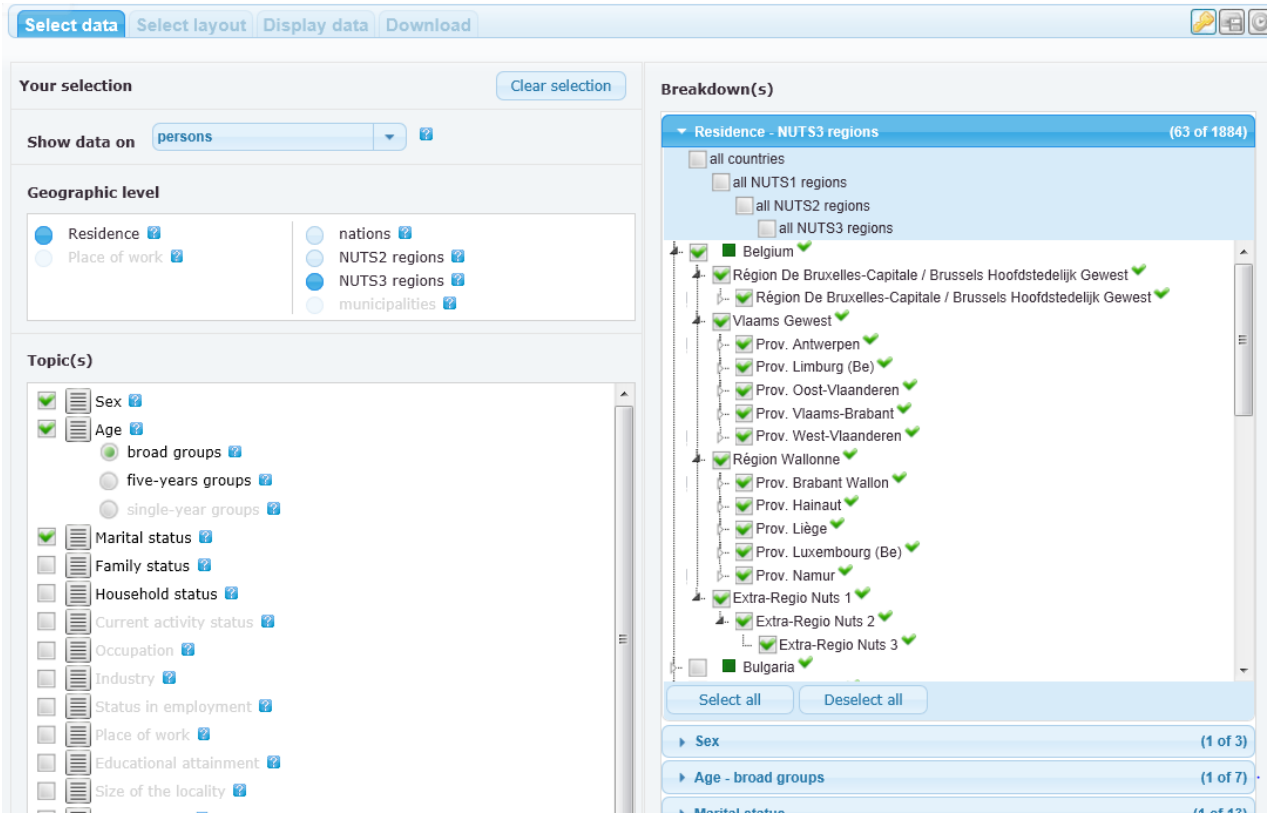
# IV. Extensions and Future Steps

## A. Improvements to the Census Hub

29. A number of improvements, mainly to the GUI of the Census Hub application, are planned in short term. For instance, the *Select dataset* and *Specify dimensions* screens will be merged. The selection of the hypercube will be transparent to the end-user. Based on the user's criteria and a weighting algorithm proposing the most suitable hypercube for a given selection, as well as performance considerations, the system will make the appropriate choice.

30. Further enhancements will enable the user to perform cumbersome operations with as few clicks as possible; for example, it will be possible to select all geographical units for a single country with just one click.

31. Special attention will be paid to illustration issues. For example, red/green indicators suggesting country node unavailability/availability will be changed to appropriately-shaped indicators, so that they are not misleading to people unable to distinguish between these colours.



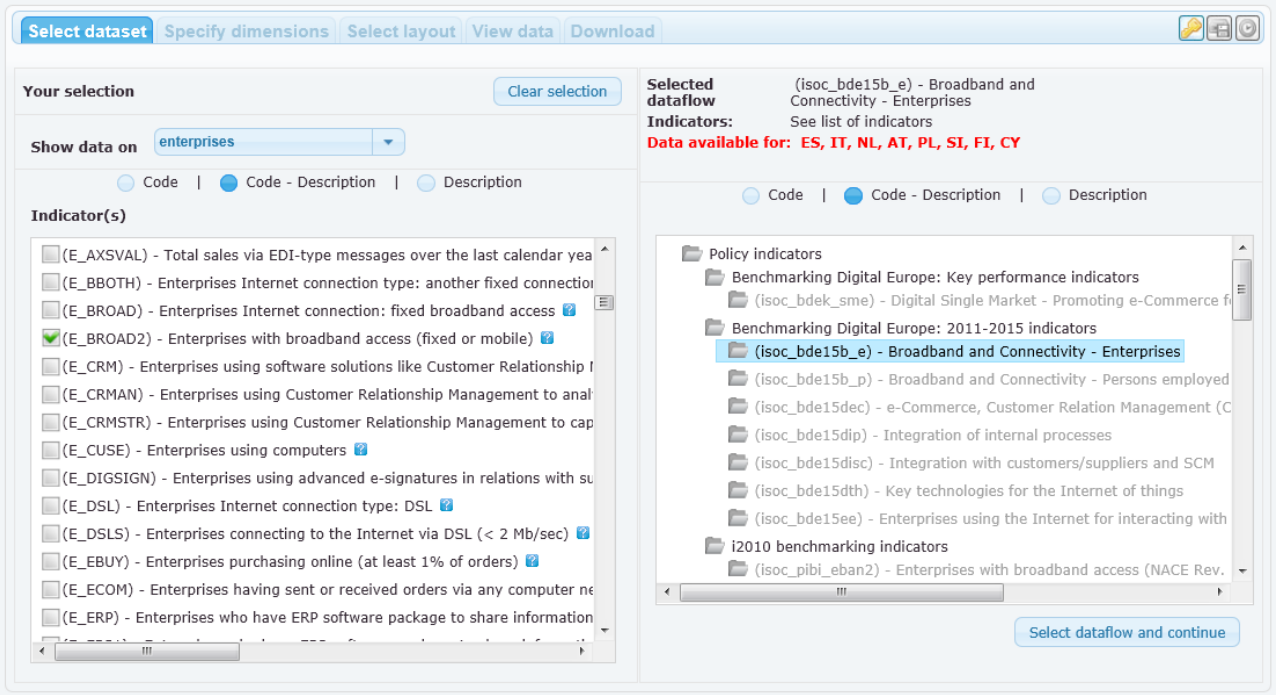

## B. The ICT Hub

32. The first attempt to reuse the experience obtained and the tools developed in the context of the Census Hub is with the 'Project on a Future European Infrastructure for Exchange and Dissemination of ICT statistics' (also referred to as the ESS.VIP.BUS ICT project or simply ICT project). The main objective of the ICT project is to modernise and standardise the production process for ICT statistics by building on the accumulated experience of other business and technology projects, such as the Census Hub and the SDMX-RI projects.

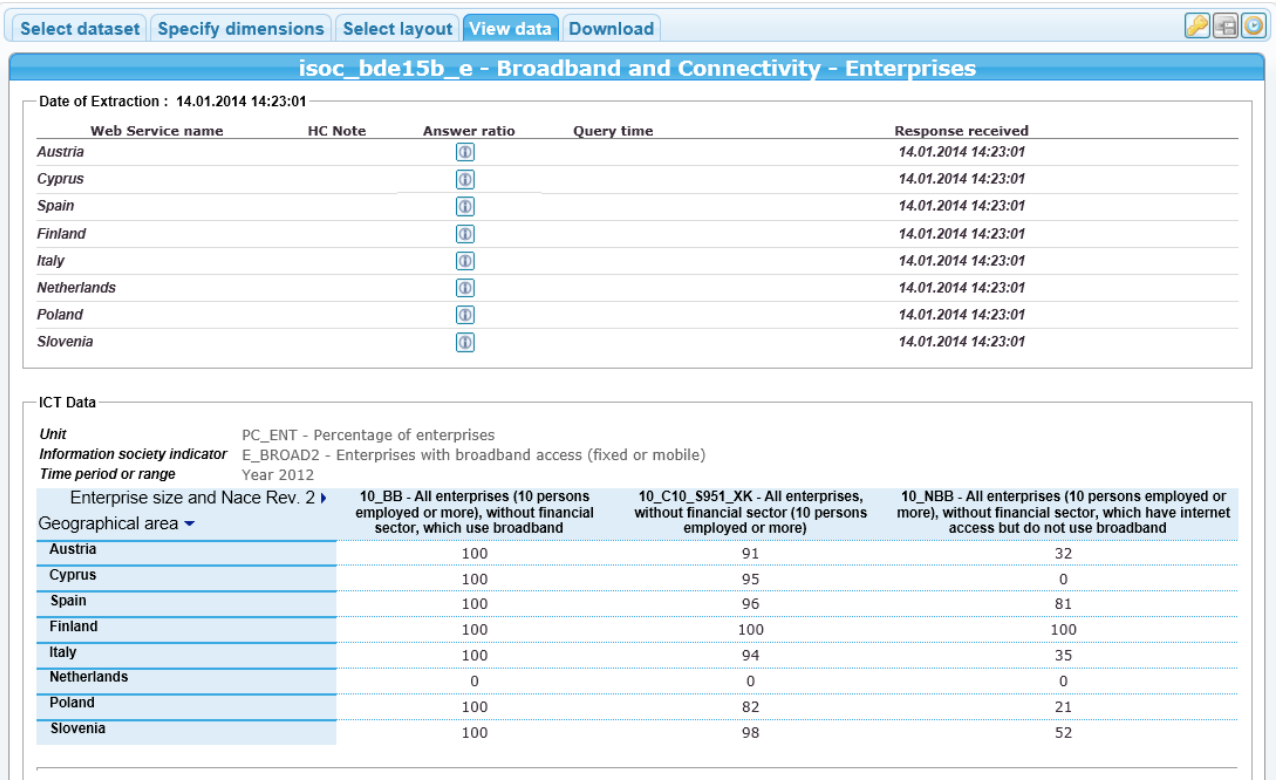
33. In 2013, Phase I of the ICT project focused, among other things, on testing data transmission modes. An ICT Hub was constructed for testing purposes, on the basis of the latest Census Hub version, and then it was appropriately extended to the domain of ICT statistics. Capitalising on the Census Hub experience, the project team achieved with minimal investment in just a few months:

(a) to reuse the Census Hub applications and tools;
(b) to carry out limited software development on the Census Hub GUI, allowing its extension to the ICT statistics domain;
(c) to provide guidelines to the 11 pilot countries (Austria, Cyprus, Denmark, Finland, Ireland, Italy, Netherlands, Poland, Slovenia, Spain, Sweden) on how to reuse the existing SDMX-RI infrastructure;
(d) to prepare test data sets to simulate dissemination data in the pilot countries' dissemination databases, which were in turn employed in relevant demonstrations and tests.

34. Following the logic of the Hub, the end-user accesses the GUI of the ICT Hub application and selects a ‘dataflow’ (‘hypercube’ under the Census Hub). The user employs the ICT ‘indicators’ (‘topics’ under the Census Hub), to locate the appropriate dataflow under the ICT category scheme:




35. After specifying search criteria, the end-user submits the query, which is in turn decomposed, transmitted to the SDMX-RI modules of the pilot countries and translated appropriately to their dissemination databases. The responses, flowing in the opposite direction, are composed and presented on the Hub.



36. Although the statistical production process is similar for different statistical domains, there can be domain-related particularities that need to be handled under a general-purpose Hub. The ICT project, besides extending the Census Hub to the ICT domain, investigates the possibility of integrating into the Hub extended functionalities that may be crucial to other statistical domains. One such functionality is the execution of 'aggregations on-the-fly'.

37. In the traditional way of handling European statistics, data are collected at a central location (Eurostat) where, among other things, aggregates at European level are computed. In contrast, within the Hub logic, each country node contains data relevant solely to that country and EU aggregates would need to be computed centrally. In the ideal case, countries own their data and can thus modify them at any time. Thus, aggregates should be computed on-the-fly on an ad hoc basis, upon receipt of a relevant end-user query, rather than be pre-calculated and stored at a central location.

38. The calculation of an EU28 aggregate on-the-fly on the ICT Hub can be problematic (or impossible) when a number of country nodes are unavailable (or respond slowly over the network), especially when the number and population of the unavailable countries would make the aggregate statistically unacceptable. Moreover, the calculation of an aggregate (an indicator or any derivative quantity in general) may require much more information than is available at the country nodes. Furthermore, when the derivative quantity is not a mere summation of values, the execution of complex calculations on-the-fly may be necessary. The ICT project is also investigating alternatives on how such challenges could be handled, both technically and methodologically.

## C. The Enhanced Hub

39. The attempt to extend the Census Hub to other statistical domains reveals various scenarios under which the Hub logic could be applicable. The idea of an Enhanced Hub that would support different modes of operation could be further developed, covering different scenarios:

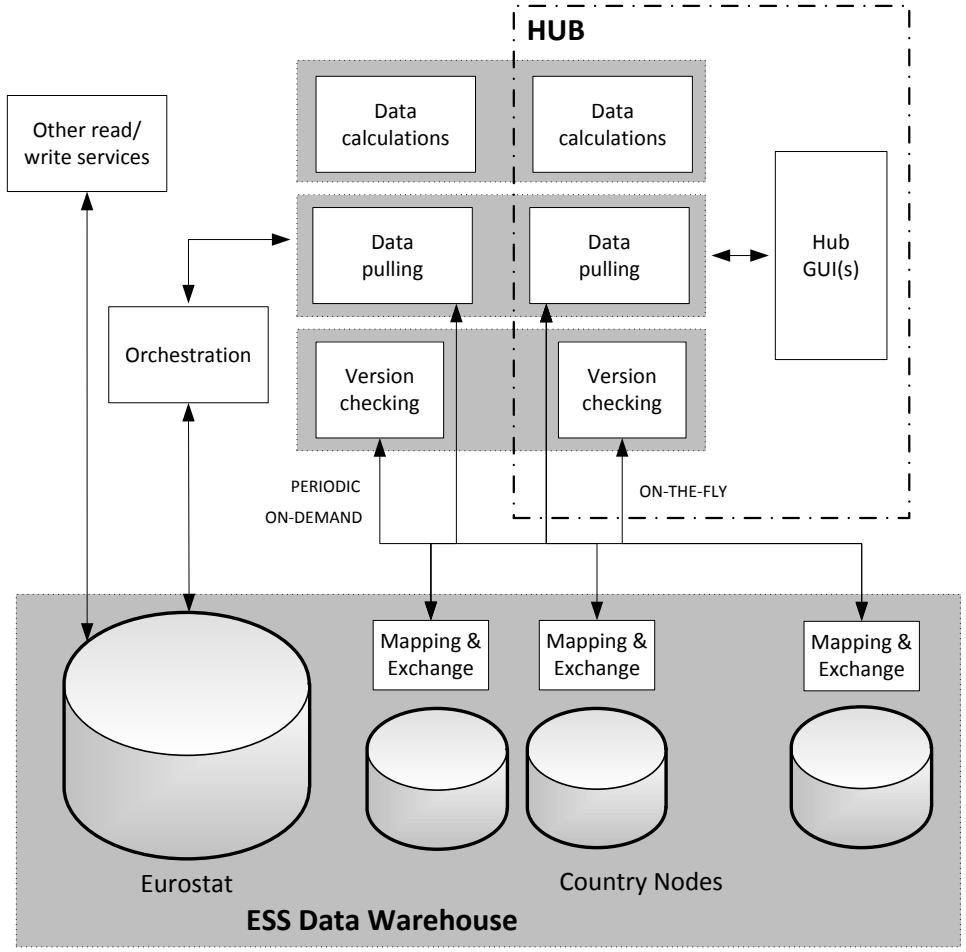
(a) **Scenario 1:** End-user requests to the Hub are answered only on the basis of country nodes currently available (no data pulling to central location). This scenario is applicable for statistical domains that demonstrate voluminous data collections, limited need for aggregations, limited data revisions and low survey frequency (suitable e.g. for population census and similar domains).

(b) **Scenario 2:** End-user requests are answered on the basis of (i) country nodes currently available and (ii) data previously pulled for unavailable country nodes. This scenario introduces the idea of periodic data pulling for statistical domains with less voluminous data collections and frequent data revisions.

(c) **Scenario 3:** End-user requests are answered on the basis of data pulled periodically (data versions between Hub and country nodes are compared). This scenario introduces the idea of consistent data versioning.

(d) **Scenario 4:** A variation of Scenario 3 where data calculations (e.g. aggregations) are executed and the results are stored after every data pull (suitable e.g. for ICT and similar domains). The scenario introduces the idea of storing pre-calculated derivative quantities after frequent data pulls for smaller or relatively static collections.

40. Such an Enhanced Hub would contain mechanisms for data pulling, data version checking and data calculations, invoked as necessary by a central orchestration module. These mechanisms could work offline (e.g. for periodic data pulling) or online (e.g. for aggregations on-the-fly in response to an end-user query on a Hub).




# V. Conclusions

41. The European Census Hub is expected to be officially announced in the summer of 2014. At the moment, the Census Hub application is online in the production environment at the premises of the European Commission and successful data exchange has been established with the countries. By the time the Census Hub is officially announced, all 28 EU Member States and 4 EFTA countries will have connected their endpoints to the Hub and will have confirmed that their data are disseminated properly.

42. The European Census Hub is already a successful European project for, among other things:

(a) Demonstrating how a multiannual persistent endeavour managed to accomplish one of the greatest dissemination undertakings ever in the ESS;
(b) Focusing on collaborative effort, with 32 countries participating actively both in terms of ideas exchange and in terms of actual effort (in the statistical and the IT context);
(c) Exploring new ways of information exchange, data dissemination and collaboration;
(d) Having provided best practices and lessons learned on data sharing and software sharing within a large community.

43. The Census Hub project is expected to evolve through future enhancements, as well as possible extensions to other statistical domains (such as the ICT Hub) and multi-domain or multi-purpose Hubs (such as the Enhanced Hub). Dissemination of European statistics could gradually evolve into a data sharing system, in which each party owns, validates, shares and continuously maintains its information.