UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

STATISTICAL OFFICE OF THE EUROPEAN UNION

Joint UNECE/Eurostat Expert Group Meeting on Register-Based Censuses
(The Hague, The Netherlands, 10-11 May 2010)

Working paper 4
3 May 2010

Topic (ii) of the provisional agenda

QUALITY OF REGISTERS

Quality Assessment for register–based Statistics in Austria

Note by Austria[*]

ABSTRACT

Due to the transition from a conservative census in 2001 to a register–based census in 2011, Statistics Austria is facing new challenges concerning data collection, data editing, quality management and documentation. Unlike in some Nordic countries the transition period from a conservative to an administrative census is very short. Accomplishing a census test in 2006 we are currently preparing for the register based census 2011, taking a special focus on quality issues. Therefore we have started a project in cooperation with WU Vienna. The aim of the project is to establish a system in which we gain qualitative as well as quantitative assessments on census topics. In the first place we develop a set of criteria for quality measurement. In order to arrive at these quality indicators we set up a process oriented framework including different hyperdimensions that will cover different aspects of quality issues. The paper aims to describe the general idea of this framework and gives a brief outlook to upcoming milestones of this project.

[*] Prepared by Reinhard Fiedler[*], Eliane Schwerer[*], Statistics Austria and Christopher Berka[**], Mathias Moser[**], Stefan Humer[**], Vienna University of Economics and Business (WU Vienna)

I. INTRODUCTION

1. Besides Denmark, Finland, Sweden, Iceland and Norway, Austria is one of only six countries that will carry out a fully register–based census in 2011. Unlike in many of these countries the transition time is very short. For example, it took about 20 years to switch from a traditional census to a fully register–based census in Finland. In the interim period the traditional census was gradually substituted by administrative data and there was enough time to enhance data quality of data sources by intensive cooperation between the National Statistical Institutes and the data owners. Before the register–based census test in 2006 we had only little experience in the use of administrative data. Both Statistics Austria and the data owners were facing a new situation. Cooperation with many data owners had to be established or existing cooperation had to be intensified. To assess the quality of data sources an accompanying survey sample with approximately 20 000 people was carried out, but it was just aimed for the census test, not for the actual census in 2011. Thus after we had accomplished the census test we faced the task to build up a system to assess quality based on the experiences of the census test. We initiated a project in cooperation with the WU Vienna. The aim of the project is to plan and implement a quality framework that makes it possible to asses the quality of the register–based census.

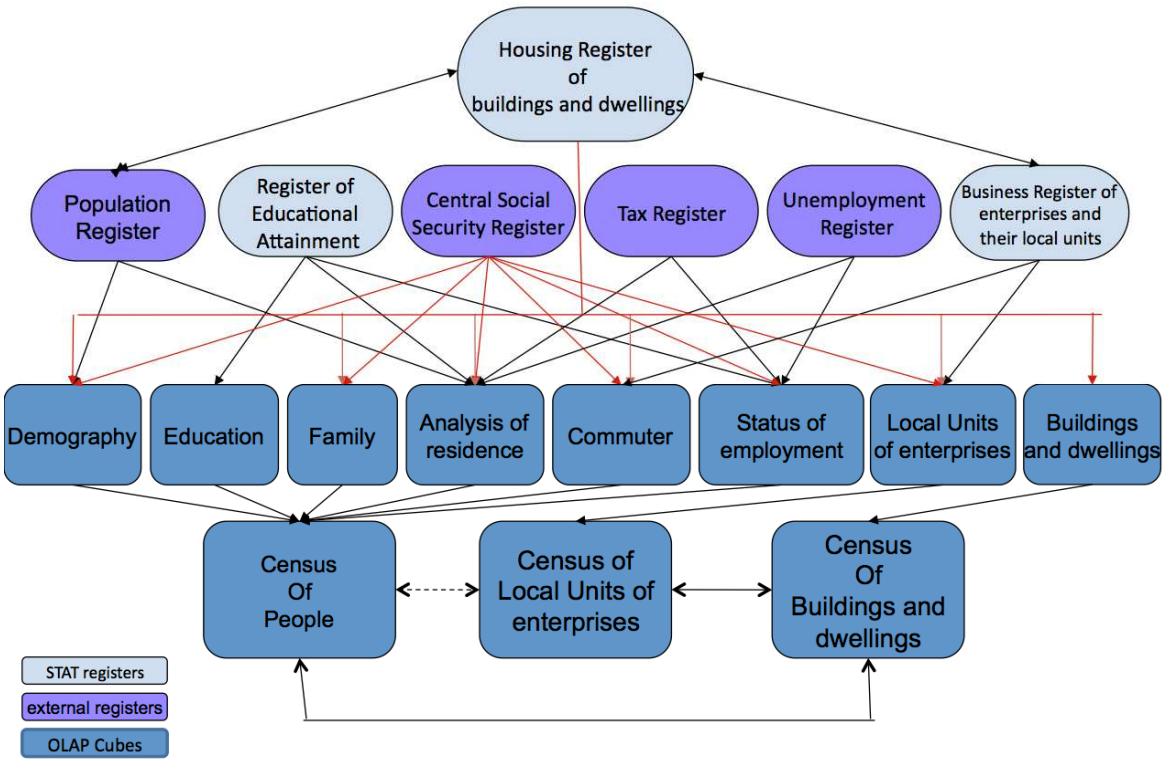
II. DATA SOURCES

2. For the census test seven base registers and additionally several comparison registers are merged. The base registers are used to determine the dimensions like the number of buildings and dwellings, the number of enterprises or the number of people with main residence in Austria. They also provide information about the core topics needed for the census. The comparison registers are mainly used for cross checks, quality issues and to complete information not or only partly included in the base registers. The base registers act as comparison registers for other attributes, too. The ”backbones” of the census are the Central Population Register (CPR) and the Central Social Security Register (CSSR). Other base registers are the Tax Register (TR), the Unemployment Register (UR), the Register of Educational Attainment (EAR), Business Register of enterprises and their local units (BR) and Housing Register of buildings and dwellings (HR). All these registers can be linked with unique keys. For individuals an artificial new identifier was introduced. The artificial identifier is derived by applying cryptographic one–way functions on the personal identification number, the source-PIN, which corresponds to the identifier of the person in the Central Register of Residents; input to these functions are the name, the date of birth, and sex of the person. The cryptographic function is customized to the government body or sector for which the data are pseudonymized; hence, the artificial identifier is called a sector–specific personal identifier and is denoted, e.g., as bPK-AS if it is calculated for Statistics Austria. Various bPKs can be calculated for a person such as bPK-health, bPK-tax, etc. Similarly, bPKs are designed and obtained for enterprises. The bPKs are calculated by the E-gov Authority established in the Federal Chancellery as a part of the Data Protection Commission (DPC). Tracing back to a certain person or enterprise via a bPK with reasonable effort is only possible for the E–gov authority.

3. Several OLAP cubes for different topics (in combination with flat files) were generated based on the DB2 database. Figure 1 gives an overview of all topics. On top one can find the base registers and their usage for each topic, e.g., there was an own cube for demographic issues using mainly data from Population Register and Central Social Security Register. In a last step the eight

main topics were merged to three final cubes representing Census of People, Census of Local Units and Census of Buildings and Dwellings.

Figure 1. Registers and Topics



III. QUALITY ISSUES IN CENSUS TEST

4. Carrying out the Census test, we may identify some causes for quality problems in the register based census:

(a) How good is the data quality of an attribute in an administrative data source? This is connected to the question for which purpose a specified attribute has to be recorded in the register. Does the administrative definition correspond to the statistical definition?

(b) For many registers there exists a residual mass of persons who cannot be linked because of missing linkage keys (bPK). Since we face the problem of imperfect linkage of registers we have implemented record linkage methods. The main attributes for our linkage procedures are sex, date of birth, and address. The most important field is the address, because it is the most reliable attribute to identify a person precisely. On the other hand there are also some shortcomings such as different notations in different registers; hence it is necessary to standardize notations. Unfortunately we have only a very few attributes for our linkage process and therefore the process itself is associated with some uncertainty.

(c) To assess data quality in the first place an accompanying survey sample was used. The detailed results of the census test including comparisons between survey and

administrative data can be found in the end report ”Probezählung 2006”[1]. In the future we will use another external data source, e.g. Labour Force Statistics.

(d) The necessity of imputing missing values influences data quality too.

5. During the census test we gathered some experience on how to assess data quality. Intensive cooperation with data owners and comparisons with surveys and other administrative sources are the key tools. Based on these experiences we set up a project ”Quality assessment in the register based census”. The project should serve the following purposes:

(a) Establish a checklist for quality evaluation on different dimensions.
(b) Establish a workflow–framework to get (quantitative) quality indicators

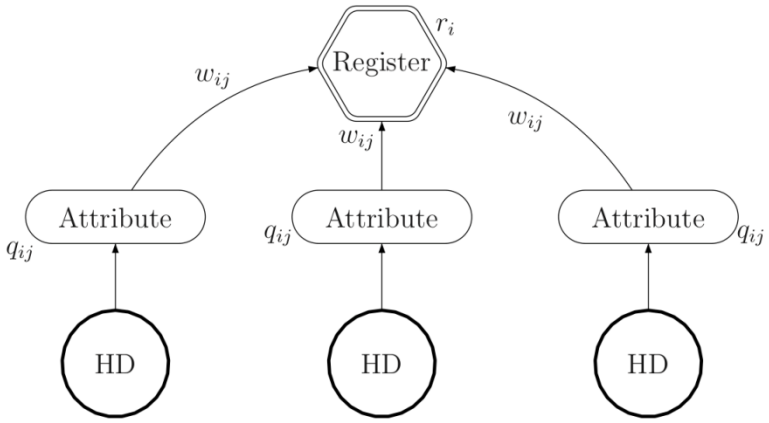
IV. A PROCESS-FLOW ORIENTED FRAMEWORK

6. As already mentioned there is some knowledge regarding the quality of different registers respectively attributes used for the census. Yet this knowledge has not been transferred into specific processes resulting in (measurable) quality indicators. For this task we set up a formal structure, based on a paper by Daas, Ossen, Vis-Visschers, and Arends-Tóth (2009). While the latter can be referred to as a more questionnaire-based approach, we found it necessary to use additional information. The main focus of this framework is the assessment of data accuracy.

A. Covering quality within different Hyperdimensions

7. Starting our quality assessment with administrative register we wish to derive quality indicators qij for each attribute in each register (See figure 2).

Figure 2 Assessment of Raw Data Quality



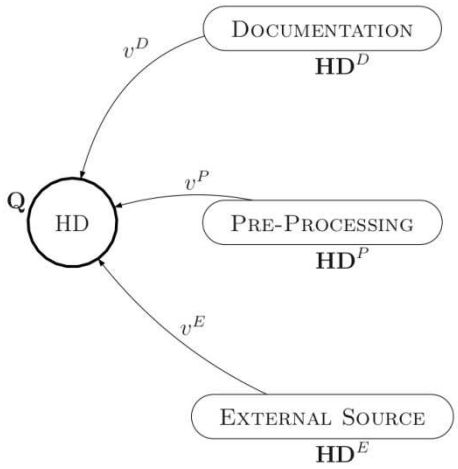
[1] See Statistik Austria (2009)

Table 1. Matrix of Quality Indicators **Q**

	Attribute 1	...	...	Attribute j
Reg 1	q_{11}	...	...	q_{1j}
...	⋮	q_{22}	⋱	⋮
...	⋮	⋱	⋱	⋮
Reg i	q_{i1}	...	...	q_{ij}

8. The quality measure for such an attribute is defined as a value between 0 and 1, whereby 1 is the best possible value. The level of this measure is derived from a set of sources (referred to as "'hyperdimensions"') which shall cover most of the information available for this attribute. Each hyperdimension delivers one (aggregated) quality indicator for each attribute. In order to combine these different quality assessments we use a weight (v) for each hyperdimension, these weights accordingly sum up to 1 (See figure 3).

Figure 3 Deriving Quality Indicators using different Hyperdimensions



9. In general we distinguish three hyperdimensions:

- **Documentation** This hyperdimension describes processes taking place before the data are actually transferred from the source (data holder) to Statistics Austria as well as the documentation of the data (metadata). In other words, the reliability of the data owner is checked. In this context this hyperdimension is built by combining the questionnaires "Source" and "Metadata" of Statistics Netherlands[2]. First of all we define the properties of an optimal administrative register. Taking this pseudo register as a benchmark, each register

[2] See Daas, Ossen, Vis-Visschers, and Arends-Tóth (2009).

used for the register–based census will be evaluated. To gather the essential information a questionnaire was set up and will be filled in by experts at Statistics Austria for each register.

- **Pre–Processing** In Pre–Processing the raw data material is evaluated. The methods used in this hyperdimension range from plausibility checks to statistical assessments such as proportions of units without unique keys. This process will be automated for each attribute.

- **External Sources** The third hyperdimension refers to the actual quality of the data itself, i.e. is the contained information exact and correct? For this purpose an external source (for the register–based census we will mainly use the Labour Force Statistics as source of comparison) has to be used in order to assess the quality of each attribute in each register. Using unique keys we are able to link the external source to the registers used for the census. Accordingly we can compare the attributes of each register with the corresponding attributes of the external source. If no external source for specific attributes exists the unit comparison is replaced by an expert interview. For this purpose the experts shall give a personal opinion about the quality of each attribute in each register. As this approach seems to be pretty subjective it is only used if and only if there is no information for a specific attribute available in an external source.

B. Deriving Quality Indicators for the Final Data Pool for the Census

10. For quality assessment the process is divided in two steps. This first step leads to the so–called Census Database Ψ, which is a result of merging registers and contains all necessary attributes for the census. In a second step we assess the quality of item imputation. Conducting this step we get the Final Data Pool Ω.

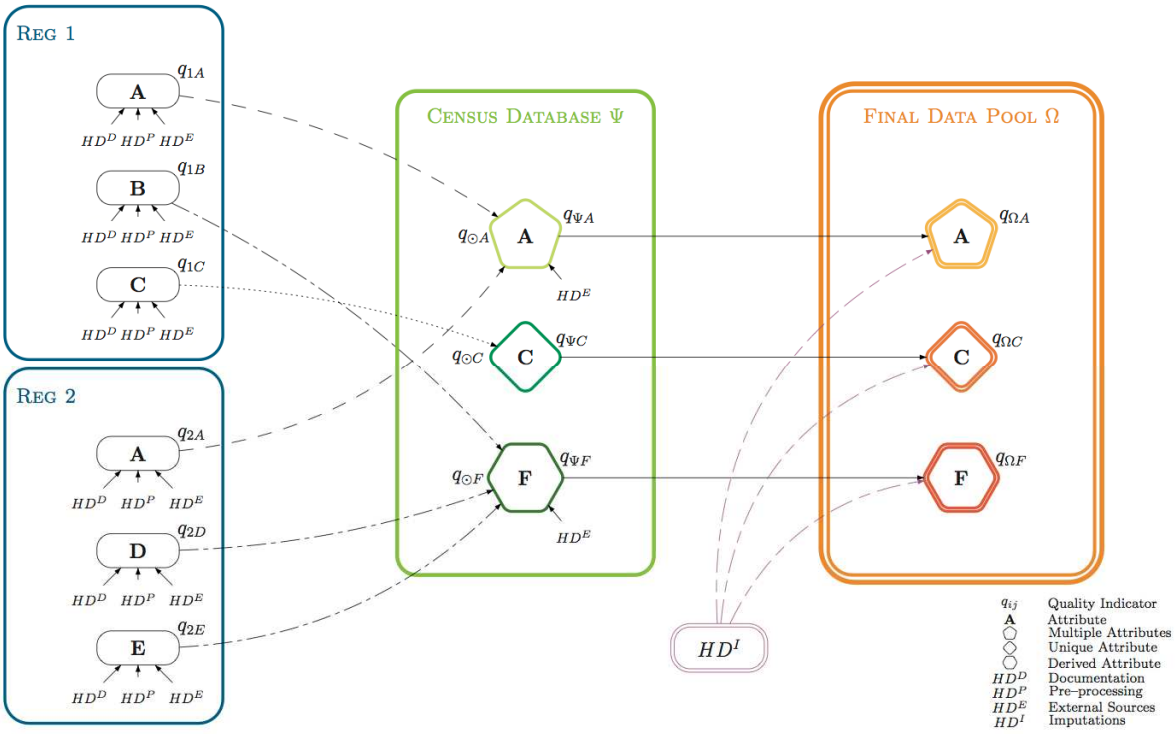
11. An overview of this process is given in figure 4.

12. **Assessment of the Census Database**: Generally we can distinguish three different cases in the process of building the Census Database:

- **Unique attributes** One attribute exists only in one register and is linked directly to the Census Database (Figure 4, Attribute **C** — e.g. *Educational Attainment*).
- **Multiple attributes** One attribute exists in more than just one register. The information is collected from these different sources (registers) and the application of decision rules (Regelwerk) finally leads to a ”valid” attribute for the Census Database (Figure 4, Attribute **A** — e.g. *Sex*)
- **Derived attributes** Different attributes are combined in order to derive a new attribute (Figure 4, Attribute **B**, **D**, **E** $\rightarrow$ **F** — e.g. *Employment Status*).

13. As unique attributes are transferred directly from the corresponding registers to the Census Data Base the quality indicator for this register (obtained from the three hyperdimensions mentioned above) can be transferred directly. The quality of multiple attributes has to be derived by combining the unique quality measures for each attribute in each register (e.g. experiments, combination rule theories, etc.). Additionally we use an external source (in this case the Labour Force Statistics) to compare the attributes included in the Census Database. This is particularly important for the evaluation of multiple and derived attributes.

Figure 4 **Quality Framework for the Austrian register–based Census**



14. **Item Imputation:** Administrative data do not always cover all attributes for each unit. Accordingly some items will be missing in the end. In order to overcome this shortcoming imputations are used to get to the Final Data Pool for the Census. Thus this hyper- dimension shall cover the influence of these processes on the quality for the attributes. Imputations are only used in the final step to get to get to this Final Data Pool.

15. **Assessment of the Final Data Pool:** The Final Data Pool consists not only of register–based data (Census Database) but also includes imputations. This remaining quantity is not yet covered within the quality assessment. We therefore need a further quality indicator which covers these additional information. Accordingly another hyperdimension is defined: Imputation (HD^I). The weight of this indicator (which shall be approximated by the proportion of imputation within the whole population) influences the overall quality indicator for each attribute within the Final Data Pool, i.e. if the proportion of imputation for attribute G is higher than for attribute H, the (negative) influence on the overall quality indicator for attribute G is accordingly stronger than for attribute H.

V. CONCLUSION

16. As mentioned in the introduction we are in the initial stage of our project. The first steps included collecting all information concerning quality assessment in register based data, specifying targets and setting up the process–flow oriented quality framework presented in this paper. This framework shall act as a road map for future milestones. The next important step involves establishing the checklists for quality evaluation. These checklists shall contain all information we need for Eurostat quality reporting and lead to the derivation of quality indicators q for each attribute in each register. However we have to specify, how to link these fundamental quality

indicators from the source registers to gather the corresponding quality indicator in the census database. Finally the integration of the hyperdimension Imputation will lead to the final quality indicators for the Final Data Pool. This framework represents a very general approach to assess quality of registers. Therefore it can and shall not only be applied to the census, but also to other projects using administrative data.

References

Statistik Austria (2009): *Bericht über die Probezählung 2006, Ergebnisse und Evaluierung*. Tech. rep. Statistik Austria.

Daas, P., S. **Ossen**, R. **Vis-Visschers**, and J. Arends-Tóth (2009): "Checklist for the Quality evaluation of Administrative Data Sources". In: *Statistics Netherlands Discussion Paper 09042*.

Fiedler, R. and P. **Schodl** (2008): *Data imputation and Estimation for the Austrian Register-based Census*. Tech. rep. UN/ECE Work Session on Data Editing. url: http: //www.unece.org/stats/documents/2008.04.sde.htm.

Lenk, M. (2007): "Practical Guidelines Data Integration – The Principle of Redundancy – Austrian Register Based Census". In: CENEX Paper WP2.

— (2009): "Using Administrative Data at Statistics Austria: Legal Provisions, Directors General of the National Statistical Institutes".

BGBl. I Nr. 33/2006 Registerz ahlungsgesetz.

Reiner, E. (2008): "Challenges of the Register Based Census in Austria with Special Focus on Effort and Impact of Including Small Register Bases". Workshop: Combination of Surveys and Administrative Data.

Main Association of Austrian Social Security Institutions (2007): "Well insured - Social Security in Austria".

Wallgren, A. and B. **Wallgren** (2007): *Register-based Statistics*. John Wiley & Sons, Ltd.

United Nations Economic Commission for Europe (2006): "Recommandations for the 2010 Censuses of Population and Housing". In: *Conference of European Statisticians*.

*****