# Use of administrative data and alternative data for census when applying modern technologies

**Janusz Dygaszewicz** Ph.D.

Statistics Poland

**Workshop on Statistical Data Editing**

(Geneva, Switzerland, 15-17 April 2020)

Presented:  September 2020

# Data collection channels in 2011 Census Round


- Including spatial data reference registers
- **CAII (CAWI)** – Computer Assisted Internet Interview
- **CATI** - Computer Assisted Telephone Interview (Call Center)
- Registered on hand-held terminals with usage GPS and GIS service
  **CAPI** - Computer Assisted Personal Interview

# CAxI method of data collection


CAxI

- CAWI - Computer Assisted Web Interview,
- CAPI - Computer Assisted Personal Interview,
- CATI - Computer Assisted Telephone Interviewing.

# Lesson learned

- The lessons learned from the census in 2011 lead to attempts to implement a full census 2021 instead of a representative survey as it was in 2011, while using data from registers as a census database and other auxiliary data sources
- It will therefore be an evolutionary improvement (next step) of the combined census.
- The results obtained will be directly and fast processed without the need to make estimates.

# Administrative registers

# Registers - data acquisition

**Data Owners:**

- Ministry of Finance,
- Ministry of Interior and Administration,
- Ministry of Justice,
- Agricultural Social Insurance Fund,
- National Health Fund,
- Agency for Restructuring and Modernisation of Agriculture,
- Agricultural and Food Quality Inspection,
- Agency for Geodesy and Cartography,
- State Fund for Rehabilitation of Disabled Persons,
- County Offices,
- Commune Offices,
- Regional Offices,
- Telcoms,
- Energy Suppliers,
- Office For Foreigners,
- Social Insurance Institution,
- Housing Managers,

- For the purposes of censuses, data from multiple sources, including administrative and non-administrative sources, are collected.
- Registries and other database systems are characterized by a wide variety and complexity resulting from the fact that they are created for different purposes and are managed by different data owners. Therefore, they also vary according to the used standards of the storage, accuracy or recording methods. The lack of uniformity exists not only between the registers but also in the data inside the registers.
- The quality of the sources and data contained has an effect on the performance of the survey. Therefore, adequate quality is a prerequisite (although not the only one) to obtain correct census results. Thus, when using administrative and non-administrative sources in research, the key elements are to identify and understand problems encountered in the data, and then unify and correct them

- The selection of the sources for the National Census 2021 was based on the experience of the previous census and the data obtained annually for statistical surveys.
- Collected resources have been systematically analyzed. There is no single indicator that assesses a register in terms of subject, object content and integration with other registers. Datasets with a minimum quality criterion are downloaded and the process of their correction and enrichment are followed.
- Cooperation with administrative data owners seems to be of key importance. In Poland it is based on:
  - long-term;
  - effective;
  - transparent;
  - based on mutual trust;
  - supported by a legal basis (Act on Statistics, Census Act).

# The usage of administrative sources during the 2011 Census

- as a direct source of data ,
- as a source of information to create a census frame,
- in addition, a source of information for :
  - imputation,
  - data estimation,
  - comparison the quality of the data.

# Advantages of using administrative sources

- Cost reduction,
- Reduction of social burden,
- Improved data security,
- Availability, on annual basis, information from register-based census,
- Availability of data from administrative registers for any level of territorial disaggregation,
- Weighting, calibration, imputation,
- Data supporting indirect estimation – modelling at the unit level,
- Efficient use of existing sources.

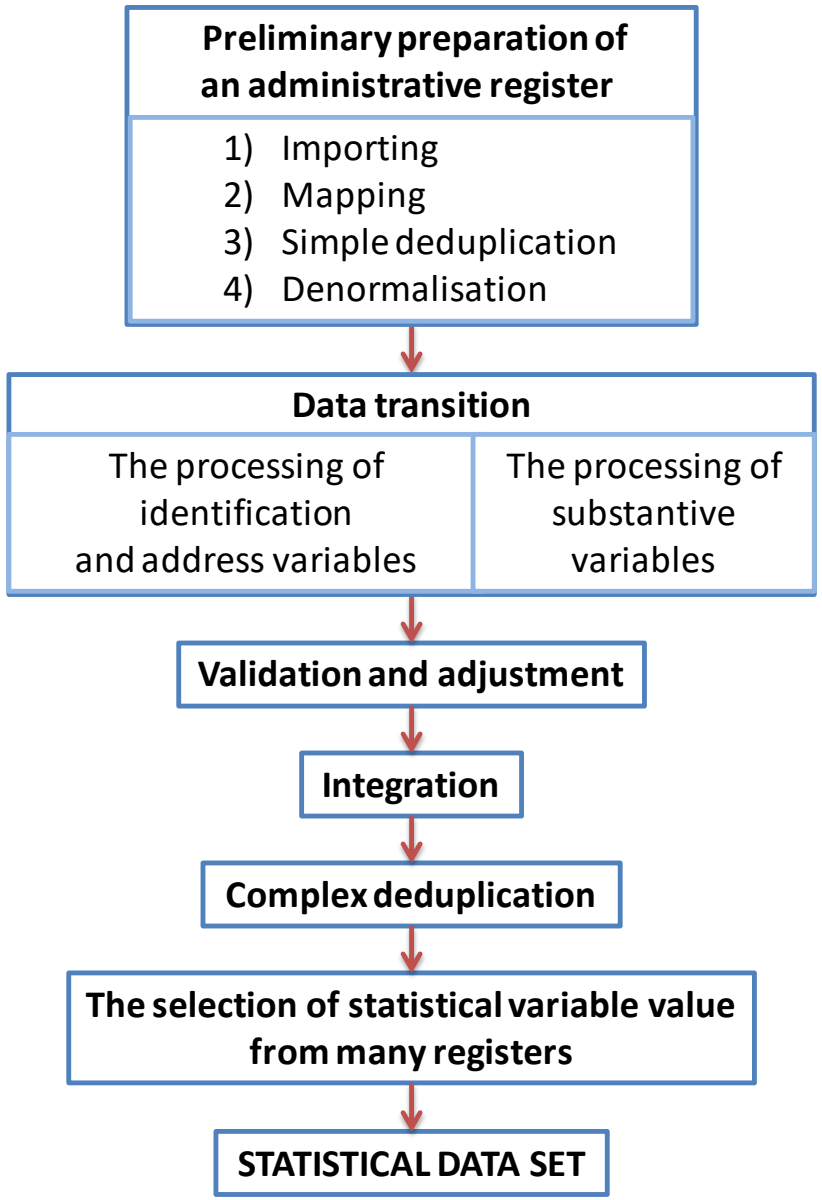
# Risks arising from the use of administrative sources

- Lack of availability of "all recommended" information,
- The dependence of official statistics on state administration bodies,
- The requirement for close cooperation between system owners and NSO,
- Changes in legislation affect the quality of the statistics obtained,
- Compliance of critical moment of census,
- Problems resulting from imputation and statistical integration of data from different sources.

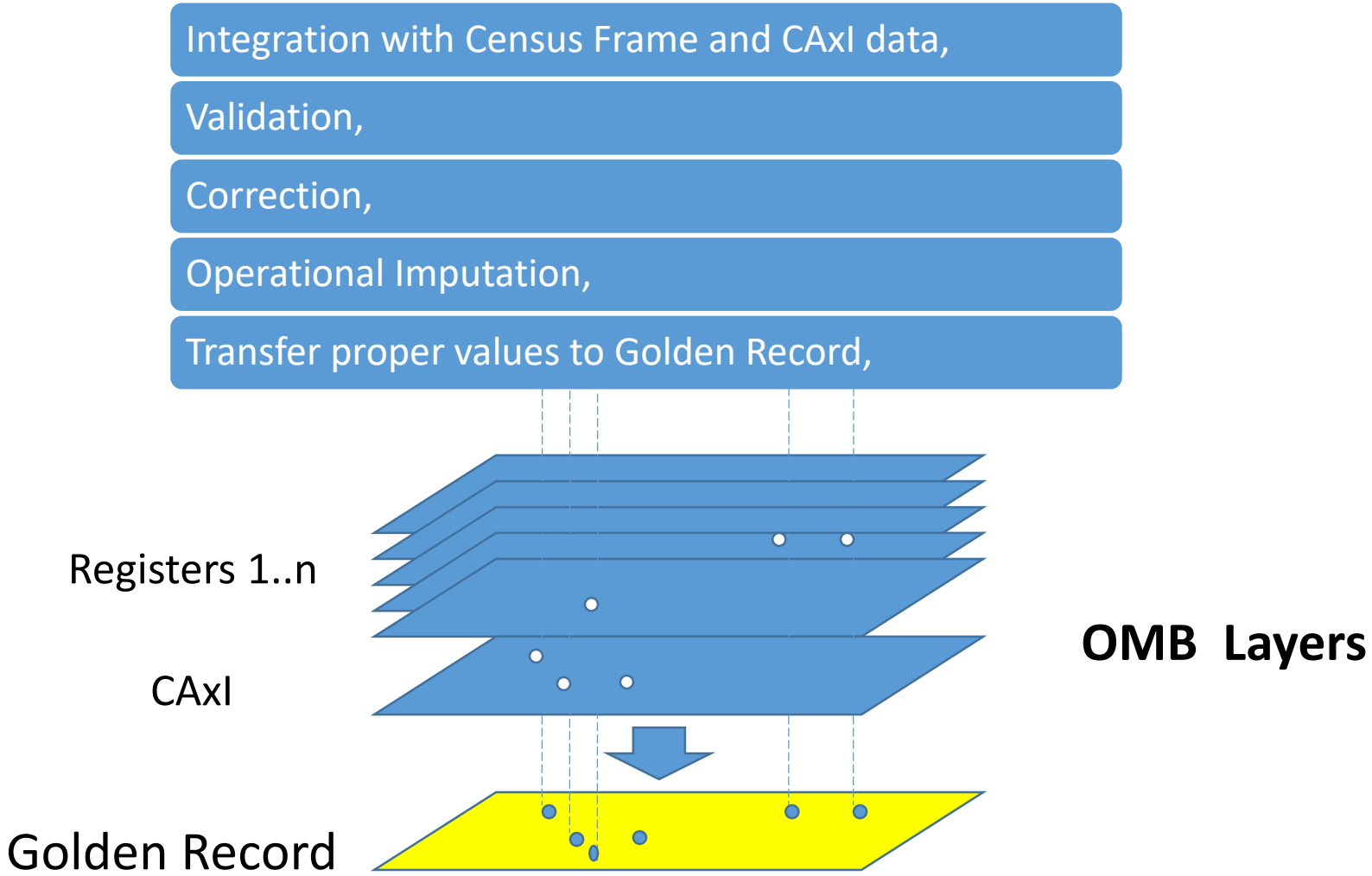
# Administrative sources

- The unit data obtained from registers are converted into statistical registers, simultaneously being subject to the process of cleaning, de-duplication and standardisation of data. The process was carried out in the DQS SAS environment. At the same time, metadata are collected on the quality of input data obtained from registers, the applied cleaning procedures and the final quality obtained after applying DQS procedures.
- The cleaned data are loaded onto the Operational Microdata Base as successive logical layers corresponding to the obtained registers.

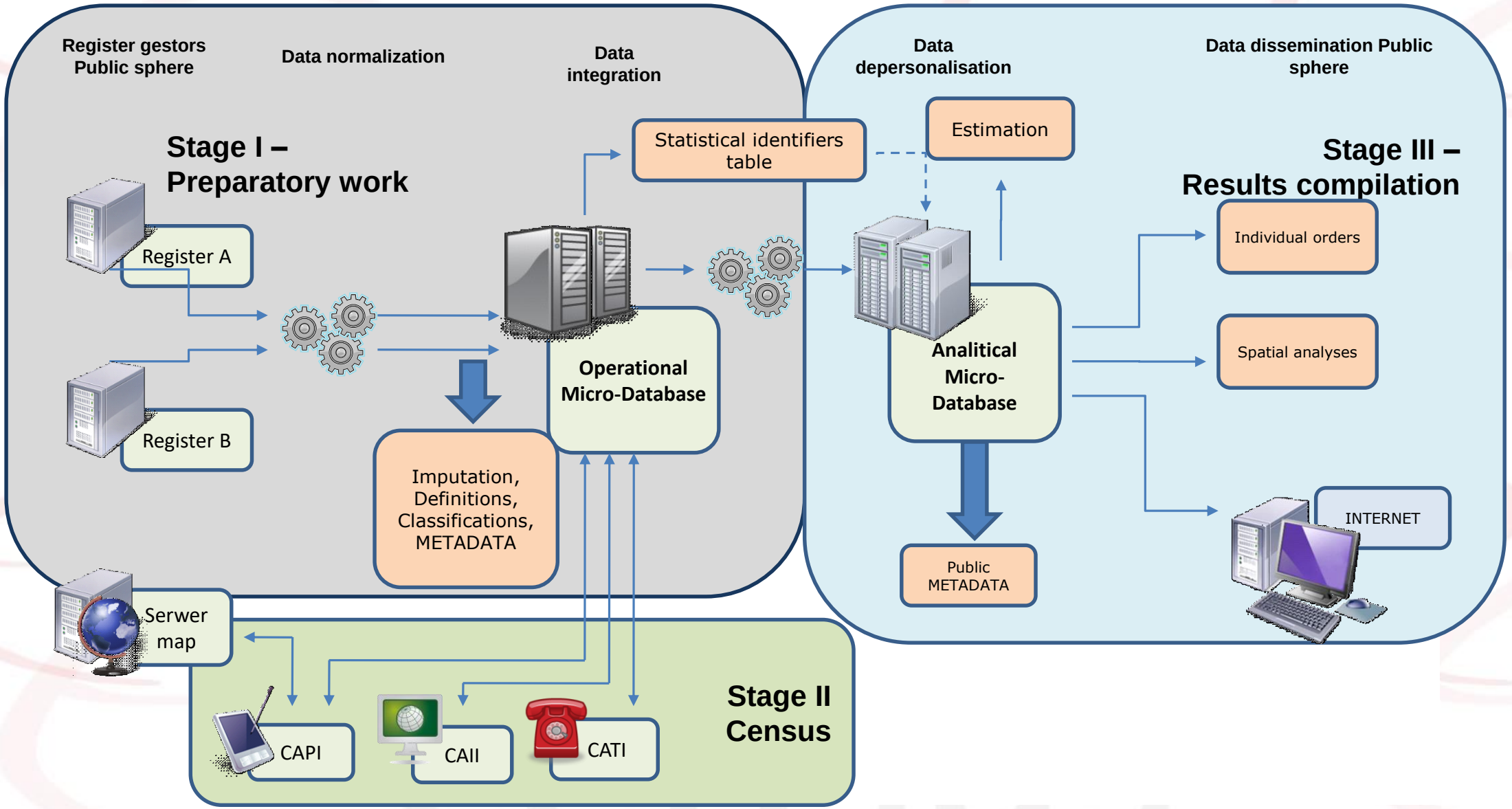
# The data transformation model from administrative sources into statistical data sets



# Golden Record generation



# Architecture solution



# The use of administrative and other data sources - metainformation

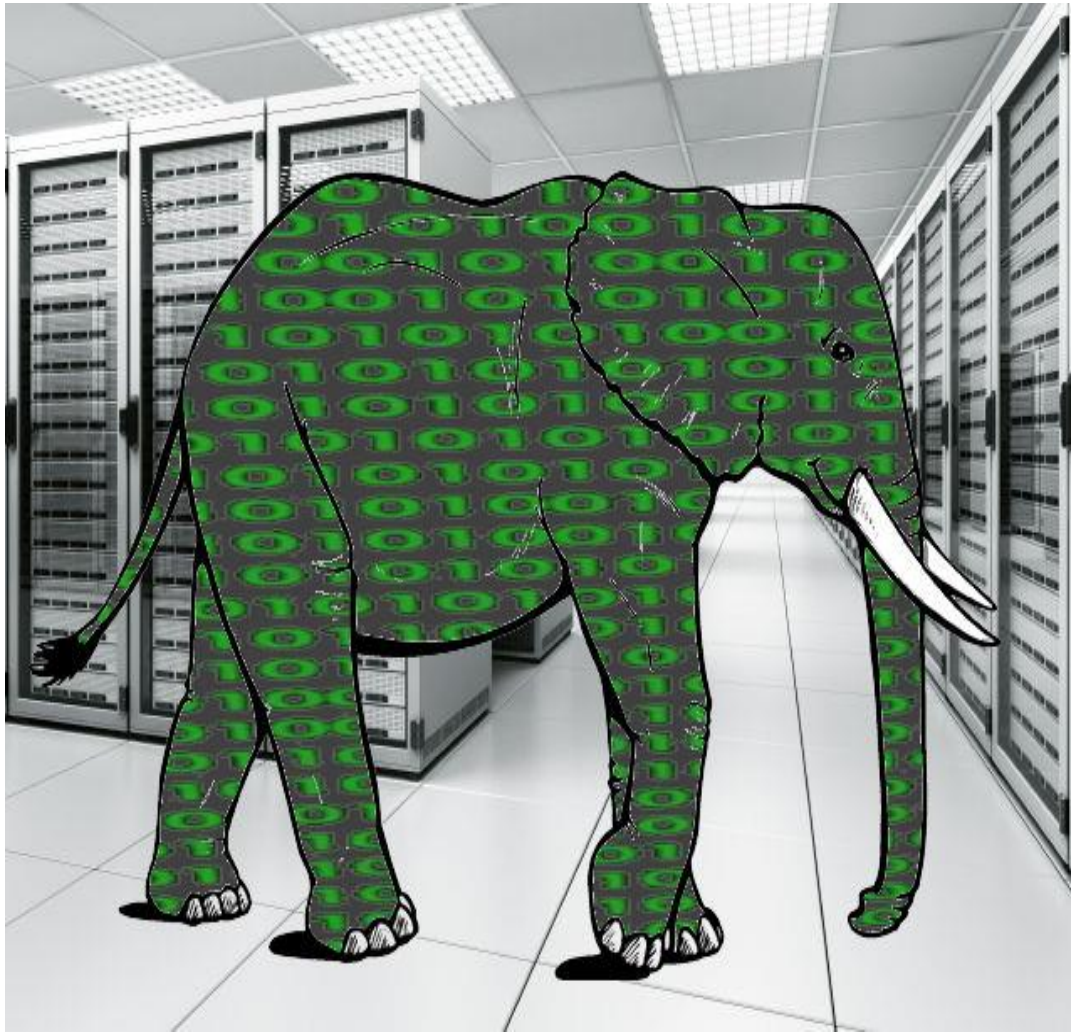
## Metainformation System includes:

- Administrative data sources description,
- Administrative data sets description,
- Rules for: integration of registers, administrative data sets transformation, developing quality indicators,
- Rules for Golden Record generation – calculating the best values of the census variables,
- Quality indicators.

# Using Big Data in Official Statistics

The use of Big Data, not only would **complement the data available to the statistics**, but in the distant future would enable the **replacement of part of the currently existing conventional surveys.**

As a result, it would enable to **reduce the burden of citizens** in filling in questionnaires **and statistical interviewers** in collecting these questionnaires, while maintaining the existing high-quality data.



# Big Data – Big Obstacles?

| | | |
|---|---|---|
| Law | Data safety | Privacy |
| Ethics | Competence | Methods |
| Technologies | Quality | Access to the data |

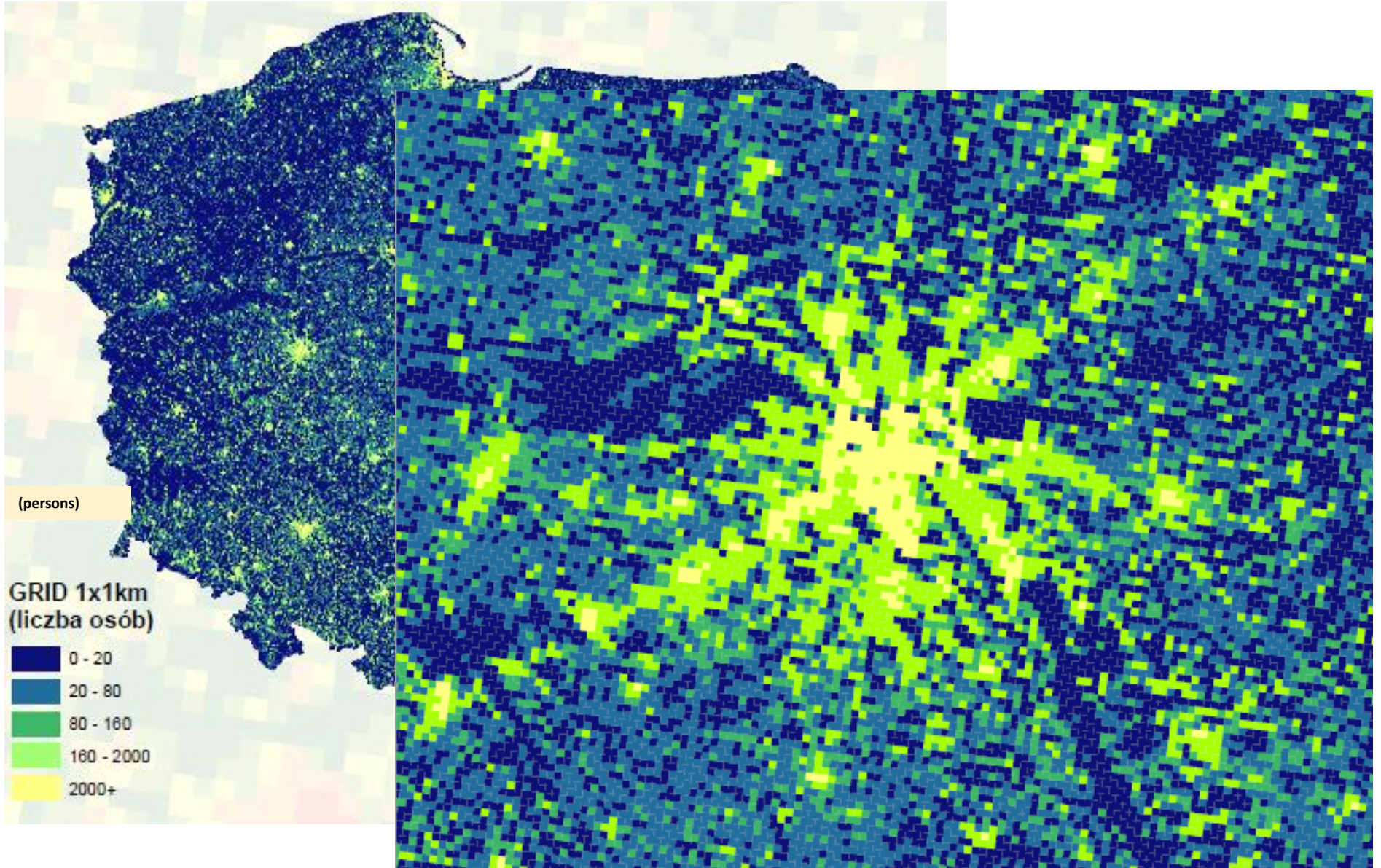
# Biga Data – Big challenges!

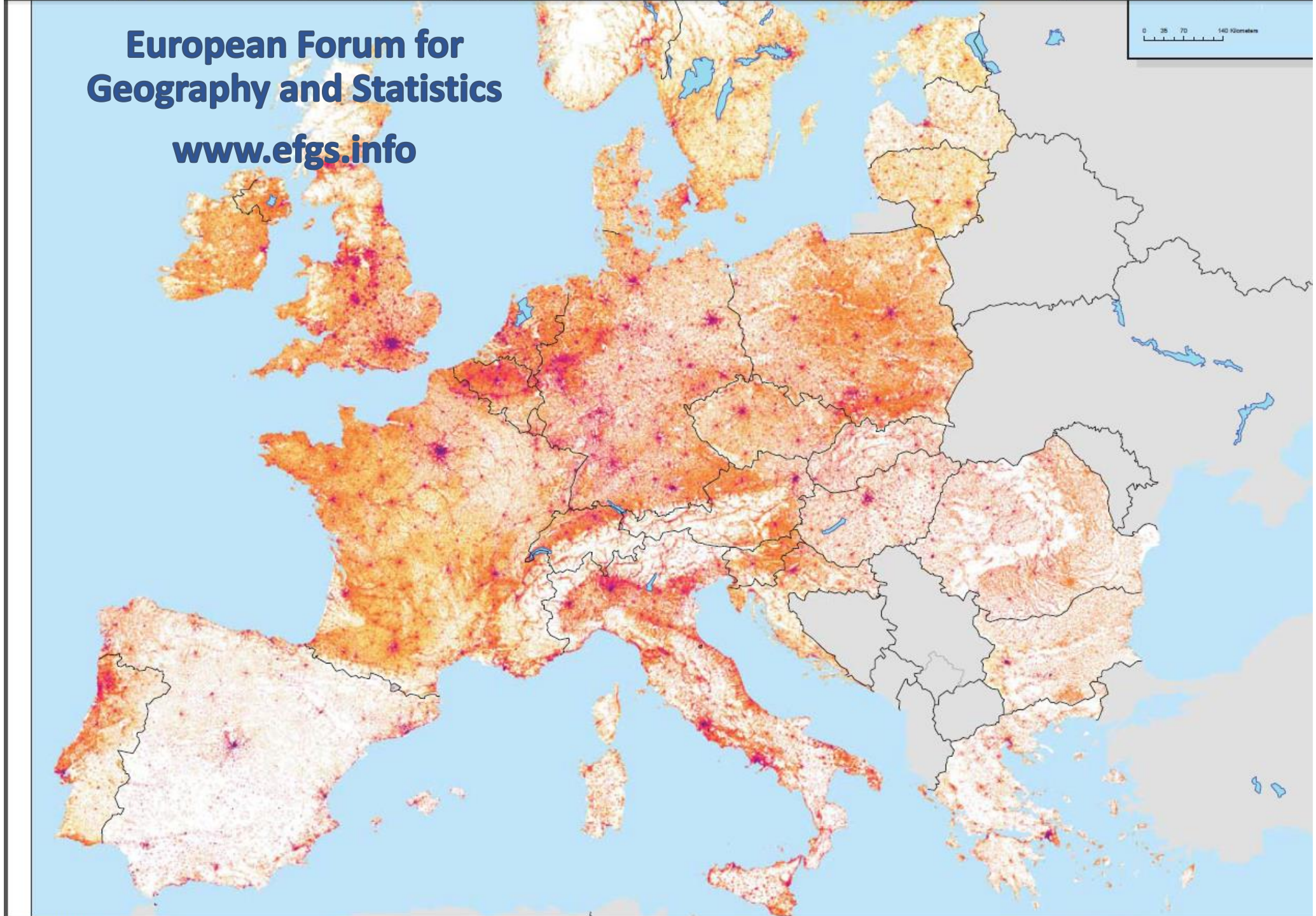
- One of the most difficult obstacles to use Big Data for the census as early as 2021 is the lack of legislation that allows collecting, analyzing and storing Big Data. This is a problem affecting all countries, although to varying degrees.
- Lack of methodology, experts and relevant legal provisions are factors preventing the use of Big Data for the census as early as 2021, but this does not mean that Big Data will not be used in subsequent editions of the census. This is a new field, but very dynamically developing. The combined forces of statisticians around the world provide a kind of guarantee that today all problems constituting a barrier will be solved in the near future.
- As part of activities after the 2021 census, it is planned to use alternative data sources, including non-administrative data, with large data volumes. Therefore, it will be necessary to use new processing methods. Currently, the possibilities of using Big Data tools in the statistical production process are being analyzed. Machine Learning components are tested as part of work on improving the quality of

# Big Data - Conclusion

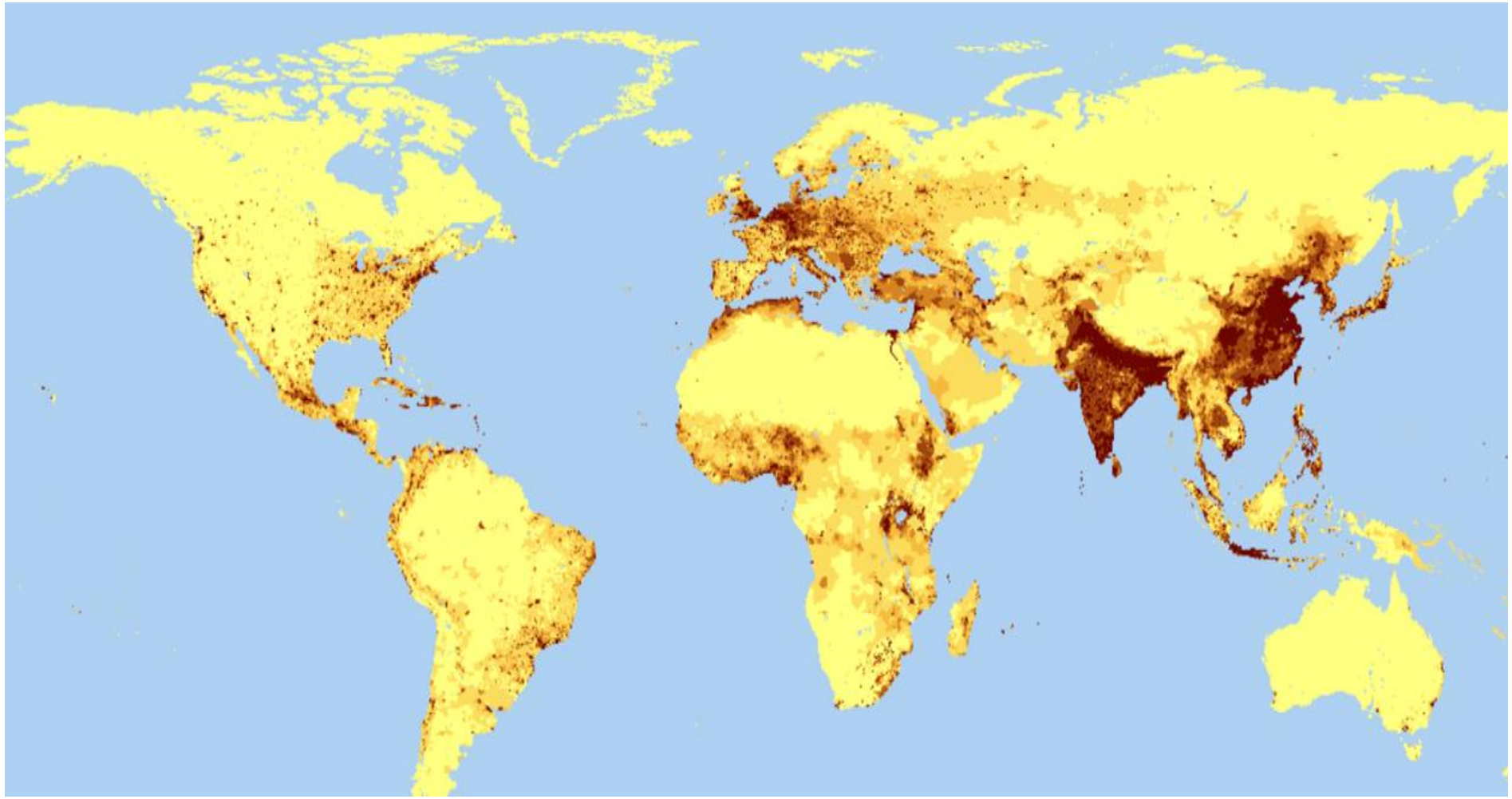
**Taking the above into consideration, the use of Big Data will not be possible in the coming census round but it will be possible after 2021 census round, including annual update in 2024.**

# Demographic data in 1 km2 grid




European Forum for
Geography and Statistics
www.efgs.info

# Global pupulation distribution



# Why applying modern technologies…

- The use of modern methods of data collection significantly speeds up their processing and development.
- Dissemination of census data in the form of micro-aggregates, aggregates, maps and result sets (e.g. hypercubes) should be implemented in an innovative way - by granting access to the analytical database by API tools.
- In the age of information society, it seems that it would be a mistake not to take advantage of the opportunity that modern technologies offer.

# Instead of a conclusion

## Census in 2002

- 180 thousands of census enumerators
- 120 mln of questionnaires

- 1 000 tons of papers

- At the end shredding census questionnaires

## Census 2011

- 18 thousands of census enumerators
- 0 questionnaires

- 0 tons of papers
- **ca. 50 mln € less**
- better data
- the more reliable results
- statistical surveys in the future

## Census 2021

**We are ready!**