UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

UNECE Statistical Data Editing Virtual Workshop 2020

(31 Aug - 4 September 2020)
https://statswiki.unece.org/x/MADUE

An overview of the editing and imputation process of the 2018 Italian Permanent census

Bianchi G., Filippini R., Lipsi R.M., Pezone A., **Scalfati F.**

ISTAT (Italy)

Outline

- **Introduction**
- **Overview of Editing and Imputation Strategy**
- **Impact of Census Innovations on E&I and Validation**
- **The Editing and Imputation System**
  - Population data process: overview of strategy and methodologies
  - Editing and Imputation management system
  - A generalized data editing for error detection
- **Concluding remarks**

Introduction

2018 NEW CENSUS

Sample of Italian households **each year**:
1,400,000 families resident in **2,800** Italian **municipalities.**

Social and **economic conditions** of the country's resident **population** at national, regional and small areas levels.

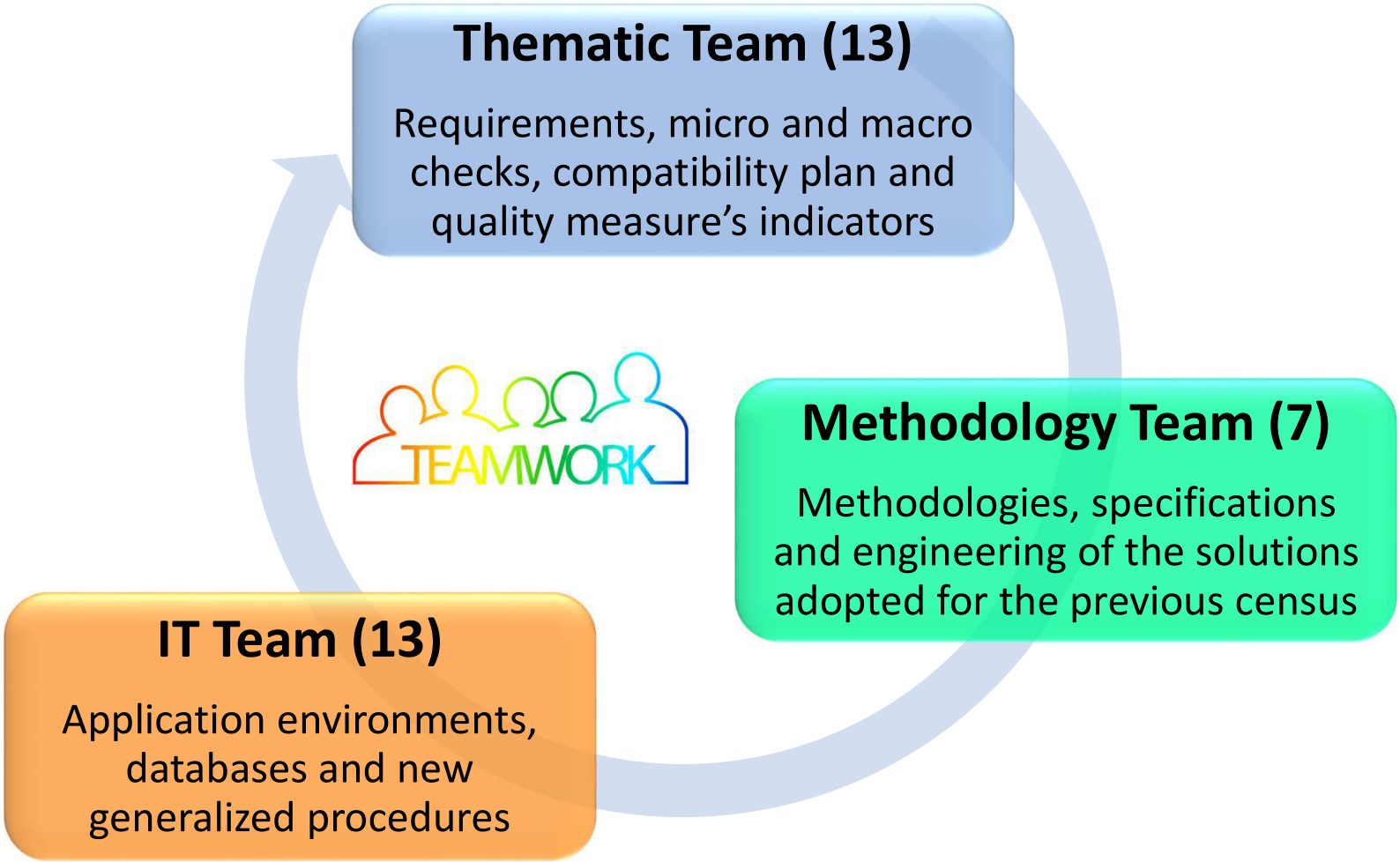
Availability of **registers**

- Local population registers (residing individuals)
- Integrative registers from auxiliary sources
- Residential address lists

Use of **web** data collection

Review of the overall
Editing and **Imputation** (E&I) **process**

The work team



Overview of the E&I strategy

Main E&I purpose: *provide a complete and consistent set of data by performing possible imputations and preserving the maximum amount of collected information*

E&I strategy: *divide the E&I problem into simpler sub-problems and find appropriate solutions for each of them*

Overall E&I process composed of several (connected) procedures addressing to specific problems and implementing suitable methods

Development and use of new techniques and software tools

Built on the useful experience of the 2011 Census, taking account of the innovations in the survey design

Impact on: E&I and Validation

Socio-economic characteristics asked on sample basis

- The reduced pool of donors for imputation of long-form variables requires careful managing of data collection and donor pool selection phases
- Sampling weights required for test, tuning and validation of E&I procedures

Availability of registers

- Improvement of the quantitative control of the forms
- Imputation of missing or inconsistent census values by matching census data and register data (**Record linkage procedure**) - reliable record identifier is required - quality of register data needs to be proved
- Imputation of missing or inconsistent census values by adding register data to census data - enlarging the donor pool

Use of web data collection

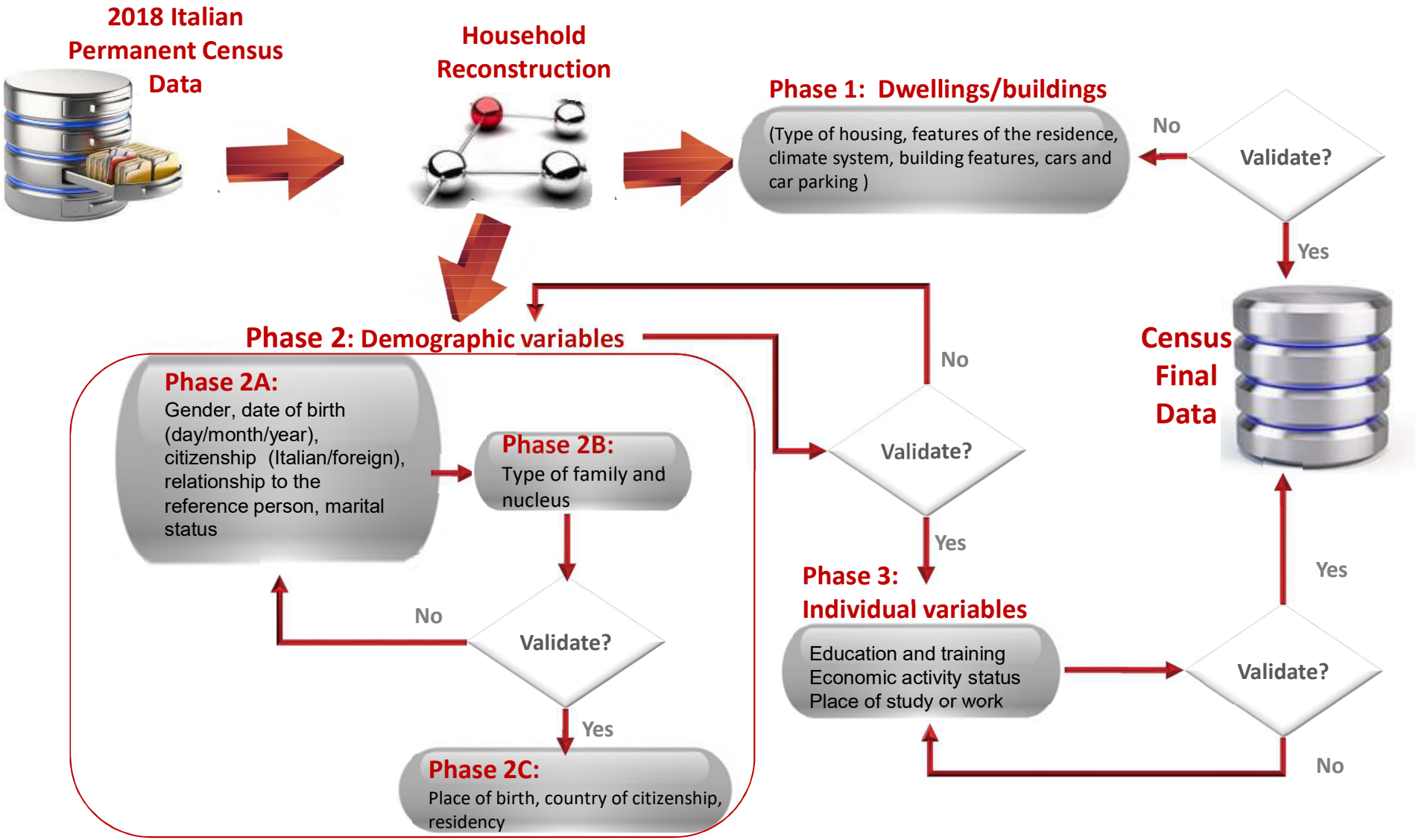
- Improvement of the collected **data quality** due to editing performed at the data capturing (**web**)

The New E&I system

Implemented for the 2018 Italian Permanent census

Main purpose	To identify and treat the non sampling errors, in order to provide a complete and consistent set of data
Quality approach	To perform different E&I tasks from data collection to the final figures
Methods	To minimize the number of changes, mainly for the treatment of not influential random errors
Monitoring	To check the main steps of E&I through a set of quality indicators
Ad-hoc documents	To assess the outcome of the procedures, focusing on data changes due to the E&I process
Subset of variables	To identify the most appropriate treatment, due to the large number of items to process and the complexity of their constraints

Flowchart of the E&I process



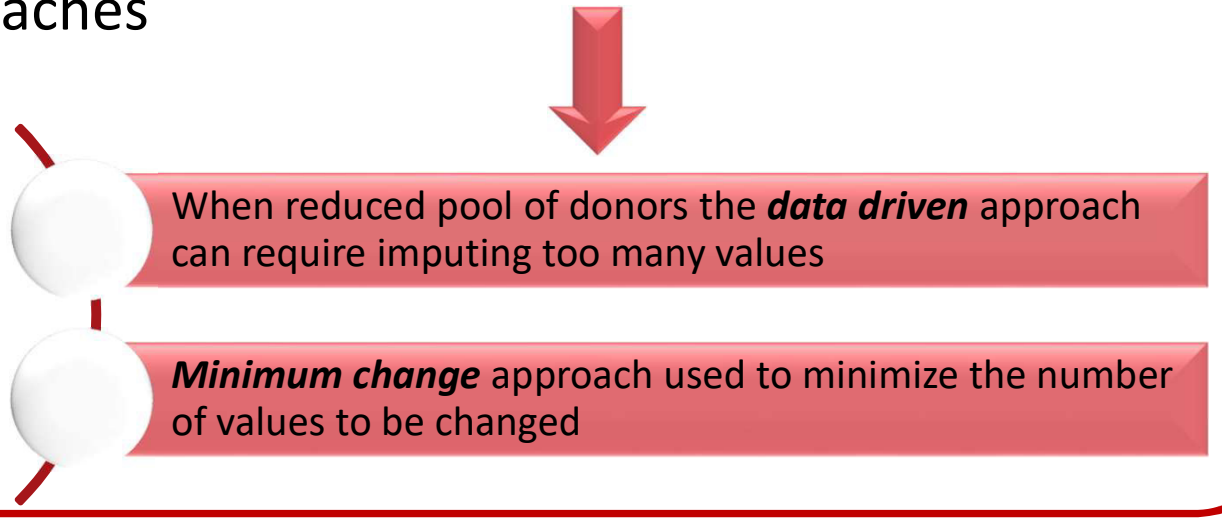
Source: Task Force C&C CP18 meeting - Rome, October 15, 2018

Main elements of the 2018 strategy (1/4)

1. Use of **D**ata **I**mputation and **E**dit **S**ystem - **I**talian **S**oftware **(DIESIS)** developed in 2001 by ISTAT and academic researchers (University of Roma “La Sapienza”).

   Based on optimization techniques, allows:

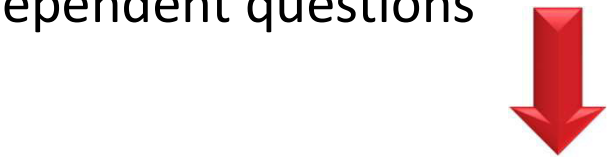
   - Treatment of qualitative and quantitative variables
   - Between-person and within-person edit rules
   - Joint use of *data driven* and *minimum change* approaches



Main elements of the 2018 strategy (2/4)

2 Identification of the respondent path

- Respondent paths used to define strata for the imputation of individual variables
- Missing responses or errors can make uncertain the identification of the right respondent path
- Automatic procedure for the identification of the most likely path based on the analysis of the responses given to filter and dependent questions



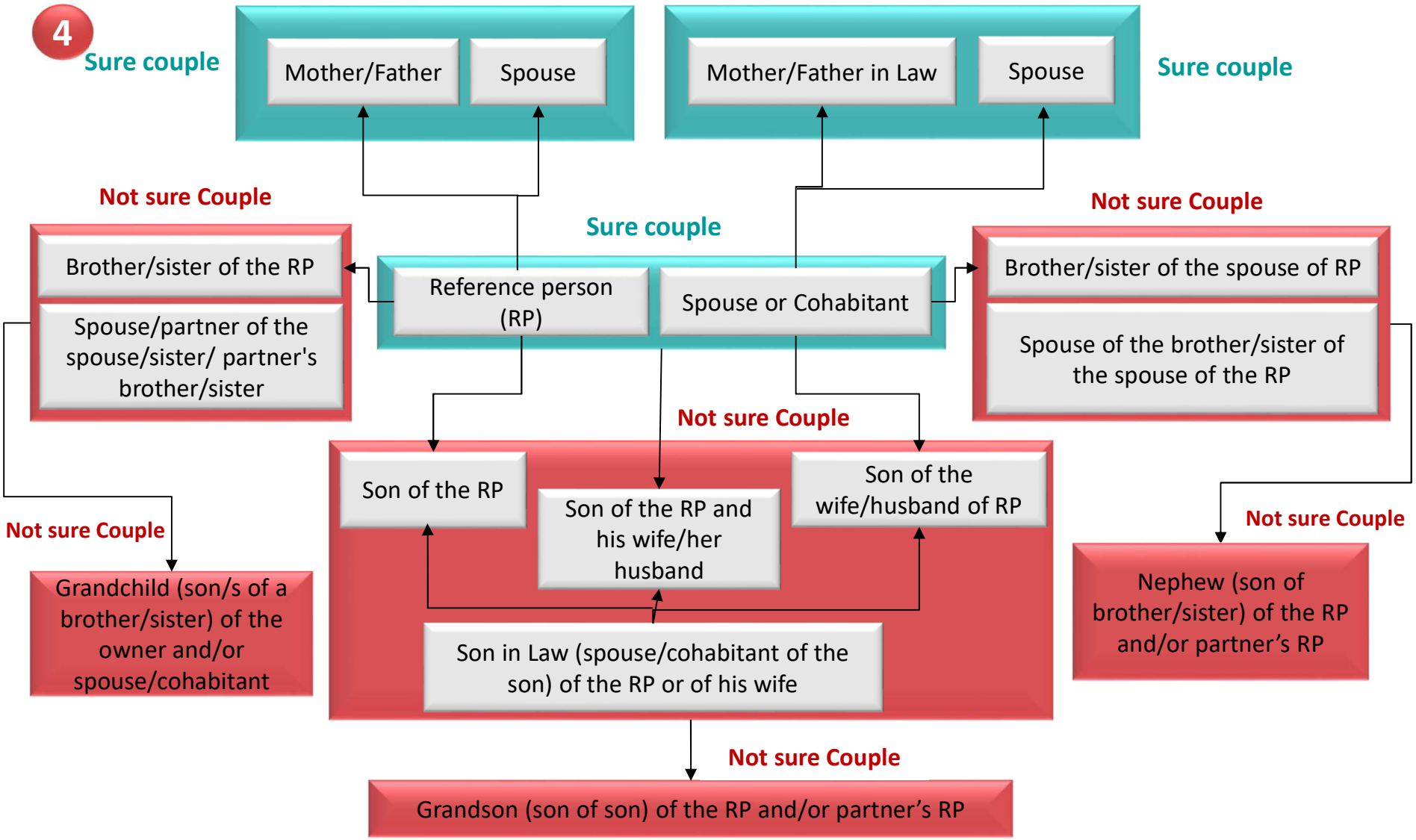
Blocks	Groups of variables involved in the same rules
Groups	Groups of blocks identified by the respondent path

Main elements of the 2018 strategy (3/4)

3 Identification of potential couples

- Based on optimization techniques implemented in the DIESIS system
- Components of couples having non-unique relationship to Person 1 (Reference Person – RP) identified prior to editing
- Score based on the responses provided to the demographic variables

Main elements of the 2018 strategy (4/4)



A generalized data editing for error detection (1/2)

1 Edits are usually represented by propositions:

- *Logical proposition* expresses a logical conditions on values of a single field
  Examples:
  (Age < 14)
  (Marital_Status = married)
- *Mathematical proposition* expresses a mathematical condition on values of at least two quantitative fields
  Example:
  (Age – Years_Married ≥14)

Logical edits are expressed only with logical propositions

Mathematical edits are expressed only with mathematical propositions

Logical-mathematical edits (or mixed edits) are expressed using both type of propositions

A generalized data editing for error detection (2/2)

2

Each rule is translated into a generalized language, created on purpose. Such language is read by the editing system.

Consistency edits are represented by the **disjunction** of two or more propositions

Conflict edits are represented by the **conjunction** of two or more propositions

Note:
- Two or more consistency edits are connected by **OR**
- Two or more conflict edits are connected by **AND**

The edit related to the **statement**:

"**If** a person's Marital status is equal to married, divorced, separated or widowed, **then** Age should be greater than or equal to 14 years"

can be written as

Consistency edit	NOT [Marital status in ('married', 'divorced', 'separated' or 'widowed')] OR Age ≥ 14
Conflict edit	[Marital status in ('married', 'divorced', 'separated' or 'widowed')] AND Age < 14

Data Editing and Imputation System (DEIS)

DEIS: generalized multiplatform application

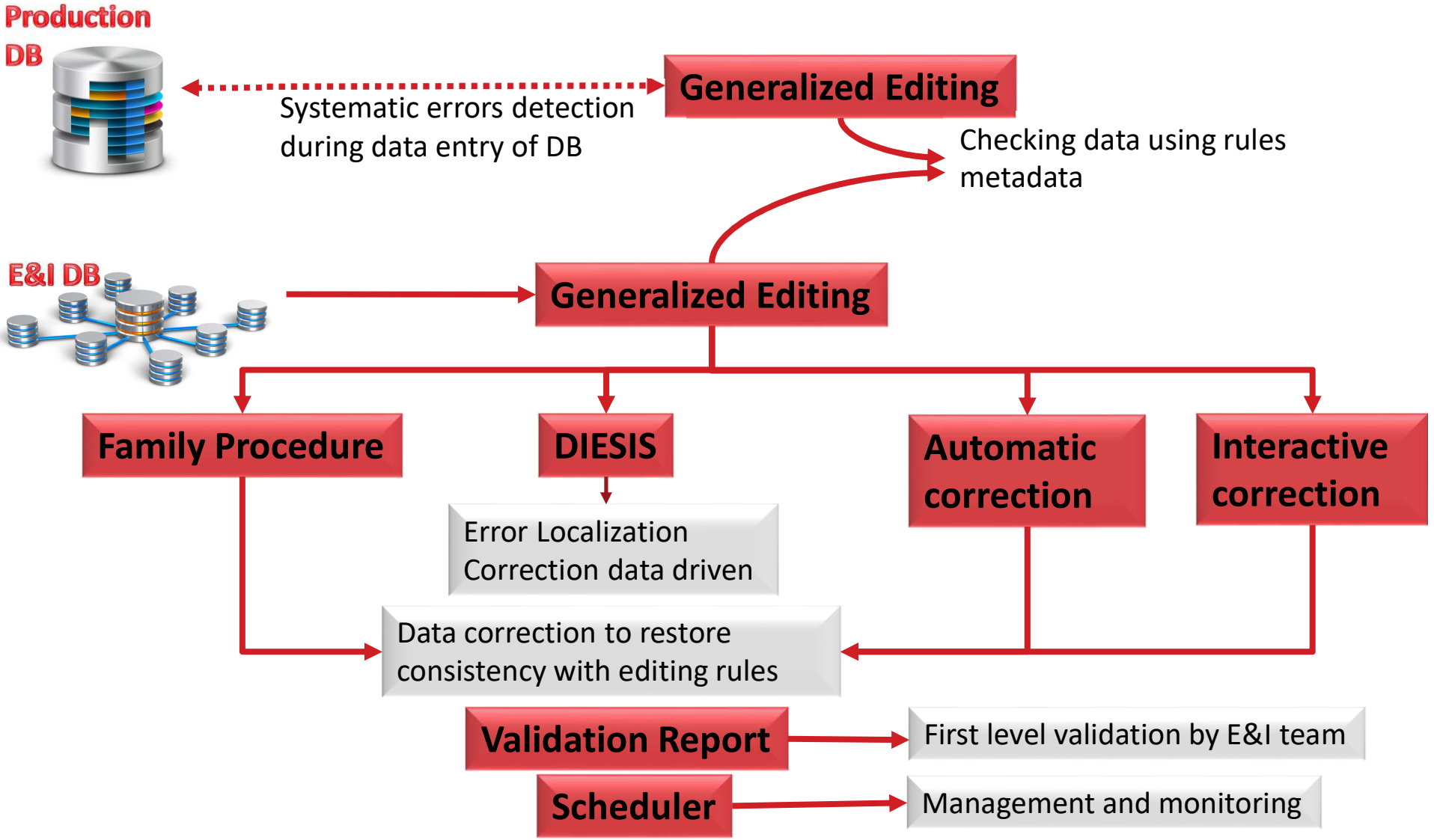
- the management
- scheduling
- monitoring of IT procedures

The Access to DEIS is regulated by user and password

Three types of users

- the **administrator**: can access every function provided by the system and create new users with different privileges;
- the **scheduler**: can start, and interrupt, the execution of the containers created and defined by the administrator and monitor the progress of each process associated with the container;
- person who **monitors**: can only check the processing status of the processes.

The E&I processing cycle



Concluding remarks

- 2018 E&I strategy based on 2011 experiences
- The new survey design aims to reduce the respondent burden but requires a **careful monitoring during production** and a more **complex E&I process**
- High efficient procedures have been developed in order to meet the timeliness requirement
- **E&I** is an achievable but **hard task**
- Generalized Data Editing and Imputation System (**DEIS**) for the management, scheduling and monitoring of IT procedures carried out for the processing of census data. This is an integrated **system of generalized services** that can be used in different context

References

- Bianchi G., A. Manzari and A. Reale, An overview of editing and imputation methods for the next Ialian censuses (Key Invited Paper), Conference of European statisticians, Work Session on Statistical Data Editing, Geneva, 13-15 May 2008.

- Bianchi G., A. Manzari, A. Pezone, A. Reale and G. Saporito, New procedures for editing and imputation of demographic variables, Conference of European statisticians, Work Session on Statistical Data Editing, Ottawa, Canada, 16-18 May 2005.

- Bruni R. and G. Bianchi, A Formal Procedure for Finding Contradictions into a Set of Rules. Applied Mathematical Sciences 6 (126), 6253-6271, 2012.

- Bruni R., A. Reale and R. Torelli, Optimization Techniques for Edit Validation and Data Imputation. In *Proceedings of Statistics Canada Symposium: Achieving Data Quality in a Statistical Agency*, Ottawa, Canada, 2001.

Contacts:
Bianchi Gianpiero (gianbia@istat.it)
Filippini Romina (filippini@istat.it)
Lipsi Rosa Maria (lipsi@istat.it)
Pezone Anna (pezone@istat.it)
Scalfati Francesco (scalfati@istat.it)