22 - 24 May 2024, **Geneva Switzerland**

UNECE

Expert Meeting on Statistical Data Collection and Sources

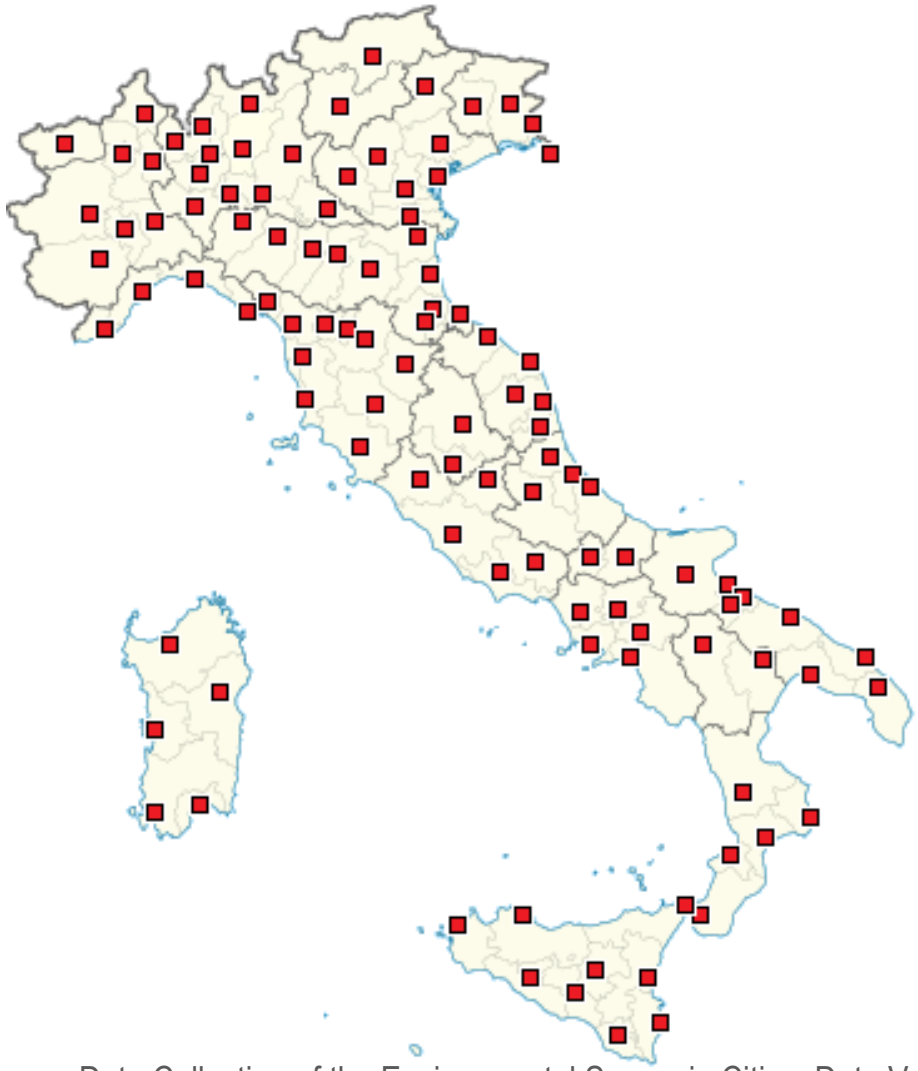
DATA COLLECTION OF THE ENVIRONMENTAL SURVEY IN CITIES: DATA VALIDATION

Domenico Adamo*, Gianpiero Bianchi**, Paolo Francescangeli**, Lucia Mongelli***
*Istat | DCAT – ATA
**Istat | DCRD – RDF
*** Istat | DCRD – RDI

Index

- Overview
- Structure
- Organization
- Data collection technique and methods
- Survey stages. Until 2022 edition
- Process innovations from ed. 2023. Validation automation
- Survey stages. After re-engineering
- Automatic validation. Some types of rules for the detection of outliers. Some examples
- A generalized data editing for error detection
- Conclusions

Overview



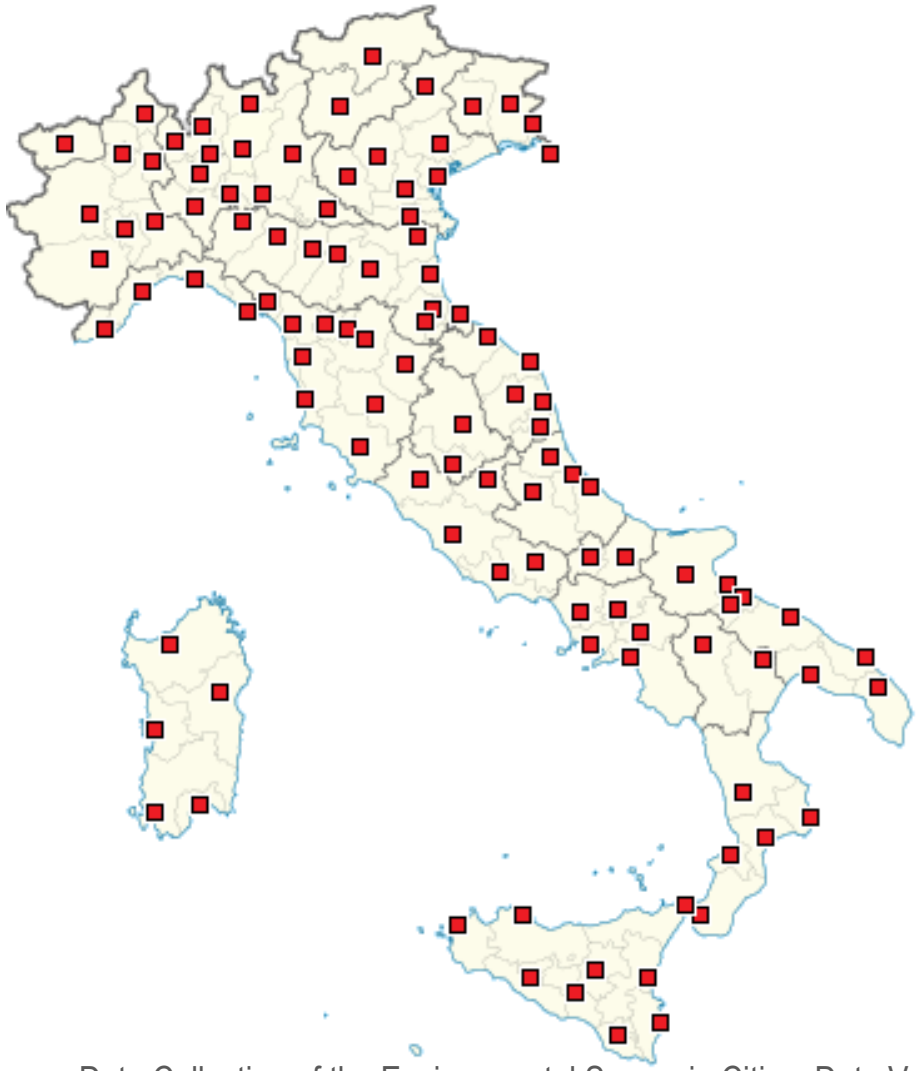
- AIR QUALITY
- MUNICIPAL WASTE
- MOBILITY
- NOISE
- ENERGY
- URBAN GREEN
- WATER
- ECO-MANAGEMENT

Urban environment data is a multiple data source statistical process, organized into **8 thematic modules**, that produces environmental indicators for **110 Italian cities** (the 109 provincial capitals +1, participating on a voluntary basis).

Most data are collected through a **CAWI survey** (Computer Assisted Web Interviewing) carried out annually by Istat since 2000. Data are collected directly from the Municipalities, via their statistical offices.

On several topics, survey data are integrated with **administrative data**, provided by other Agencies of the National statistical system (SISTAN).

Overview



AIR QUALITY

MUNICIPAL WASTE

MOBILITY

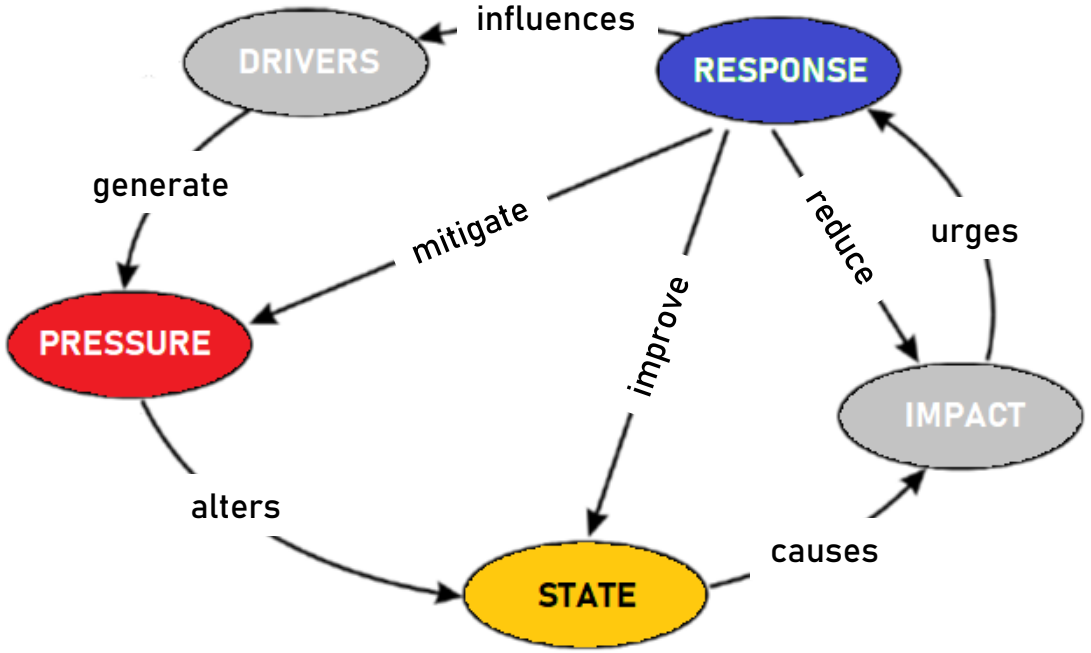
NOISE

ENERGY

URBAN GREEN

WATER

ECO-MANAGEMENT



The process provides a comprehensive information framework for monitoring the **quality of urban environment** (**state** and **pressure indicators**, according to the DPSIR model ↑), and the **environmental policies** implemented by local administrations (**response indicators**).

Overview

LEGAL BASIS

The survey on urban environment is part of the **National Statistical Programme** (PSN), managed by the SISTAN and updated every 3 years. Being included in the PSN as a **survey of public interest**, data collection is performed by law, and the response is mandatory for the reporting units.

The PSN provides also the legal basis for the **use of administrative data for statistical purposes**. Specific agreements are made between Istat and the data holders to define the features and timing of data provision, within the SISTAN regulatory framework and under the rules for the protection of personal data.

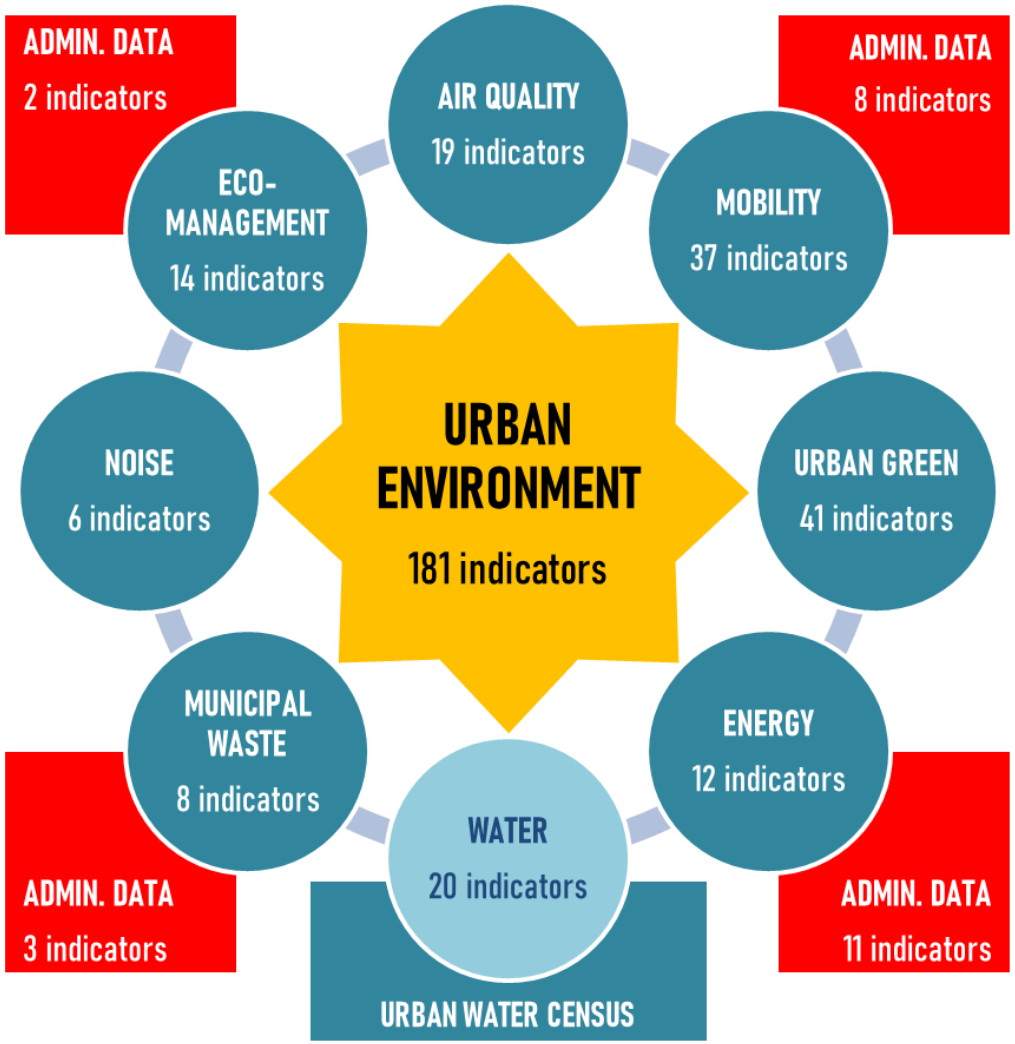
Use of administrative data allows reducing costs and the burden on respondents.

REPORTING AND ANALYSIS UNITS

The survey's reporting units are the **Municipal statistical offices**. Every year, before opening the data collection, Municipal offices appoint a coordinator and one or more reference persons for each thematic questionnaire.

Depending on the topics, reference persons collect the data directly from the Municipal administrations, or request it from other local Agencies (e.g. public transport companies). All indicators are disseminated by municipality, aggregated estimates by geographical area are also provided.

Structure



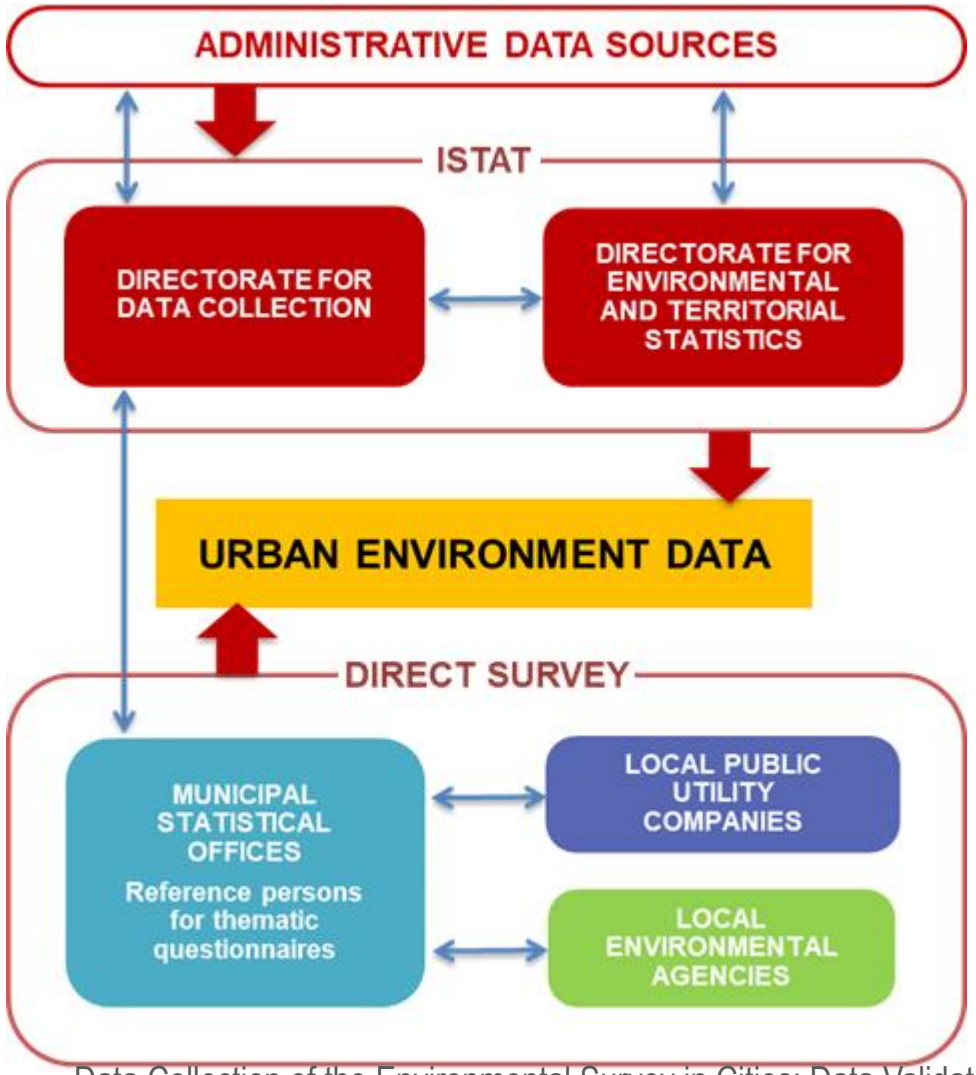
Survey data are collected through **thematic questionnaires** (except for Water, since 2018 part of the *Urban water census*, covering the whole national territory).

Four modules integrate survey data with **administrative data** on particular topics (provided separately by data holders).

Administrative data are acquired, stored and made available to statistical production units by the Data collection directorate. Being collected for other purposes, administrative data need to be processed and validated in order to produce relevant statistical information. This task is performed by the Environmental and territorial statistics directorate, which also assist Data collection in the relationship with data holders.

The entire process manages about **500 elementary variables**, to produce **181 indicators** (2020 release), 13% of which based on administrative data. Some indicators are part of the Istat set of statistical measures for monitoring the SDGs in Italy.

Organization



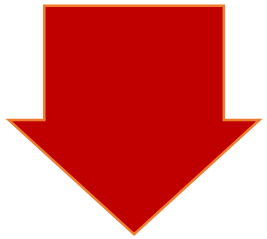
The **Directorate for data collection** is responsible for the entire process of data collection, the organization and monitoring of the survey and the acquisition of data from administrative sources.

Urban environmental data is part of the **Directorate for environmental and territorial statistics**, which is responsible for survey design, information contents, data validation and dissemination.

The wide range of topics covered makes the process of gathering information at the local level quite complex. Thus, **assistance to respondents** and a strict **monitoring of survey operations** are key factors for data quality.

Data collection: technique and methods

In accordance with the Code of Digital Administration (d.lgs 82/2005) for the collection "**Environmental data in cities**", from year 2008, a major innovation and simplification in the data collection process has been introduced.



Introduction of the **CAWI (Computer Assisted Web Interviewing)** technique for data acquisition in electronic format, through the Gino++ portal (Gathering Information Online).

Data collection: The GINO++ system

GINO++ is a generalized software, robust and flexible, allowing you to:

1. create electronic questionnaires for data collection with controlled data acquisition;
2. to set up the customized site for the investigation;
3. manage the survey (population of respondents and all actors involved, monitoring of questionnaires and registrations, contacts for reminders and reminders, reports).

Data collection: The GINO++ system

The data are collected by the Municipal Statistical Offices, which identify in the Administrations to which they belong the referents of the survey topics, who are given personal credentials to log in, enter, edit and save the data.

Through GINO++ the Municipal Statistics Offices and the contact persons of the various topics can send the data by filling in direct web questionnaires.

Automated checks are implemented to prevent the entry of inconsistent or invalid data or out of range, and the sending of questionnaires with missing responses.

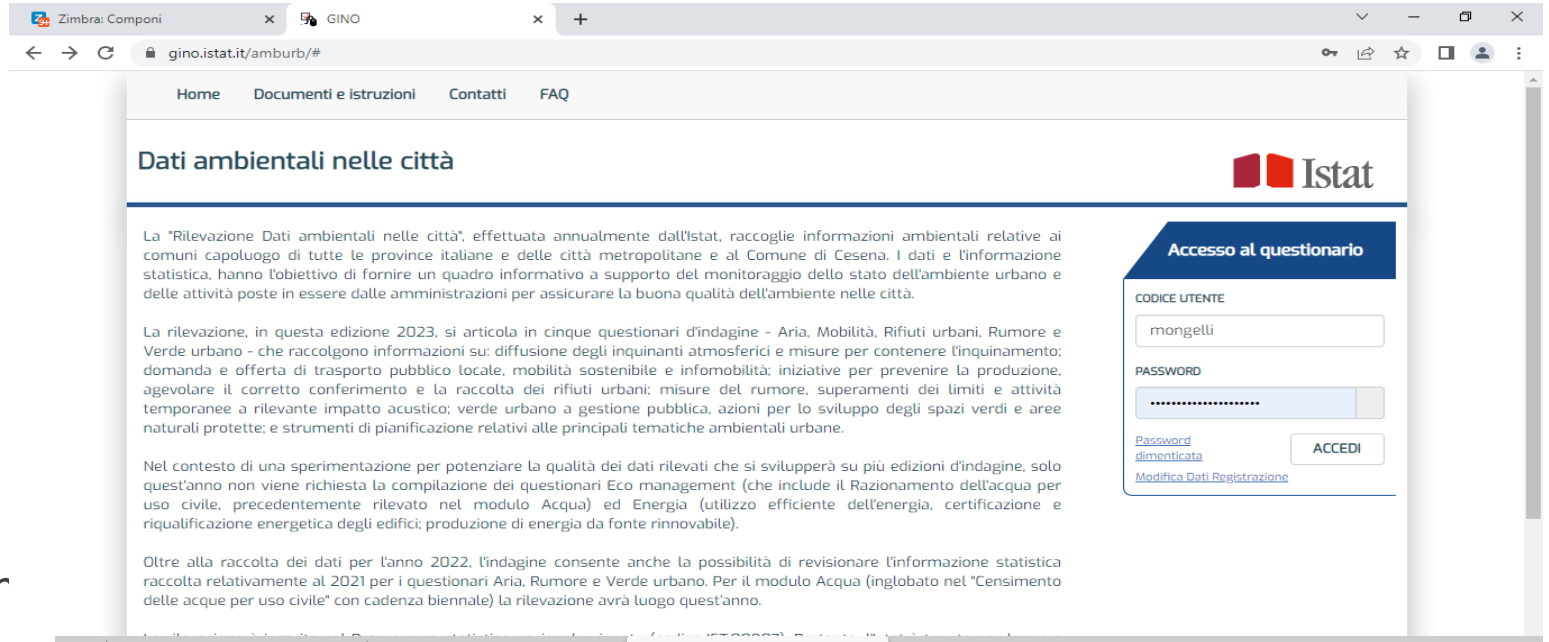
The GINO++ system

Access to the GINO++ acquisition system:

https://gino.istat.it/amburb/

In the ***Home*** are reported:

- ✓ the description of the investigation
- ✓ the obligation to reply
- ✓ instructions for access and completior the questionnaire (details in the 'Documents and instructions' section)
- ✓ the IT requirements
- ✓ the FAQ



Survey stages. Until 2022 edition

		MUNICIPAL OFFICES	STAGES	ISTAT	
				Data collection	Environmental statistics
			Survey design	Survey organization, Implementation of CAWI questionnaires	Information contents and metadata management
Questionnaire states	**Initial** - before taking over by reference person	Registration	**Data collection**	Controlled acquisition through Gino electronic questionnaire with rules	Assistance to respondents and to Data collection staff
	In process - after first opening by reference person	Data entry		Monitoring of survey operations	
	Sent - after completion by reference person				
	Checked - after preliminary check			Preliminary check and follow-up	
			Data processing		Data editing and validation
			Data dissemination		Data analysis and reporting

Process innovations from ed. 2023. Validation automation

Concerning Istat's objectives to innovate the various statistical information production processes, concerning the progressive digitalization of data collection processes, it has become necessary to design the use of innovative solutions through the reengineering of the validation phase of the questionnaires, with the implementation of further automatic control rules different from those already provided by the validity and internal consistency checks of the Gino ++ system.

The new rules would manage, at least in part, the aspects so far entrusted to the review by expert colleagues: in particular the interception of measurement errors, discontinuities of time series and other outliers.

These rules would work on a dynamic basis, by comparing the data collected with those validated in the previous edition. To be fully operational, however, this process innovation requires a testing phase, which could be concluded over the next two editions of the survey (2023 and 2024).

2023 Edition

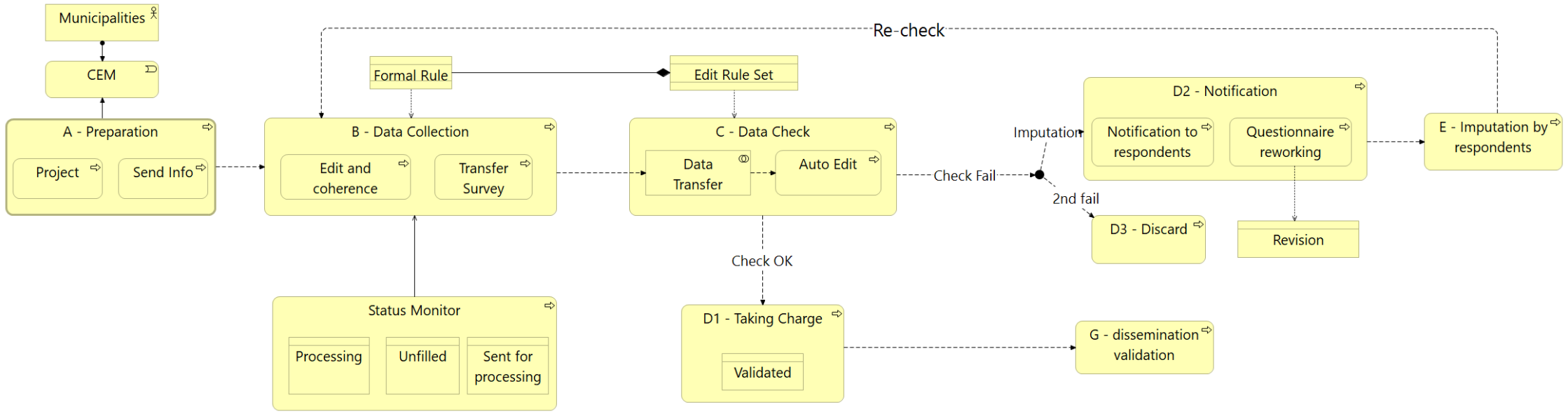
During this experimental phase, to ensure the regular conduct of the survey while maintaining the quality standard of the data collected, a diversified frequency is expected for some thematic questionnaires, which do not produce indicators intended for institutional dissemination, as in the table below:

Thematic questionnaires	2023 Edition	2024 Edition
Air	…	…
Eco-management	(suspended)	…
Energy	(suspended)	…
Mobility	…	…
Municipal waste	…	(suspended)
Noise	…	(suspended)
Urban green	…	

Survey stages. After re-engineering

		MUNICIPAL OFFICES	STAGES	ISTAT	
				Data collection	Environmental statistics
			Survey design	Survey organization, Implementation of CAWI questionnaires	Information contents and metadata management
Questionnaire states	**Initial** - before taking over by reference person	Registration	**Data collection**	Controlled acquisition through Gino electronic questionnaire with rules Automatic control of: ▪ measurement errors, ▪ historical series discontinuities ▪ other abnormal values. Monitoring of survey operations	Assistance to respondents and to Data collection staff
	In process - after first opening by reference person	Data entry			
	Sent - after completion by reference person				
	Checked - after preliminary check			Automated return on respondents	
			Data processing		Data editing and validation
			Data dissemination		Data analysis and reporting

Process Outline



- Process Spans into several Phases
  - Survey Project and Preparation (A)
  - Data Collection (B)
  - Data Check (C - D - E)
  - Data transfer to archieves and to production units (G)

- A Monitor control the data flow and records questionnaires **status** and overall completion status
- Questionnaires check can fail only once. Unfilled and double failing units are separated and stored
- Validated Questionnaires are sent out for dissemination

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

Edits are usually represented by propositions:

- *Logical proposition* expresses a logical conditions on values of a single field

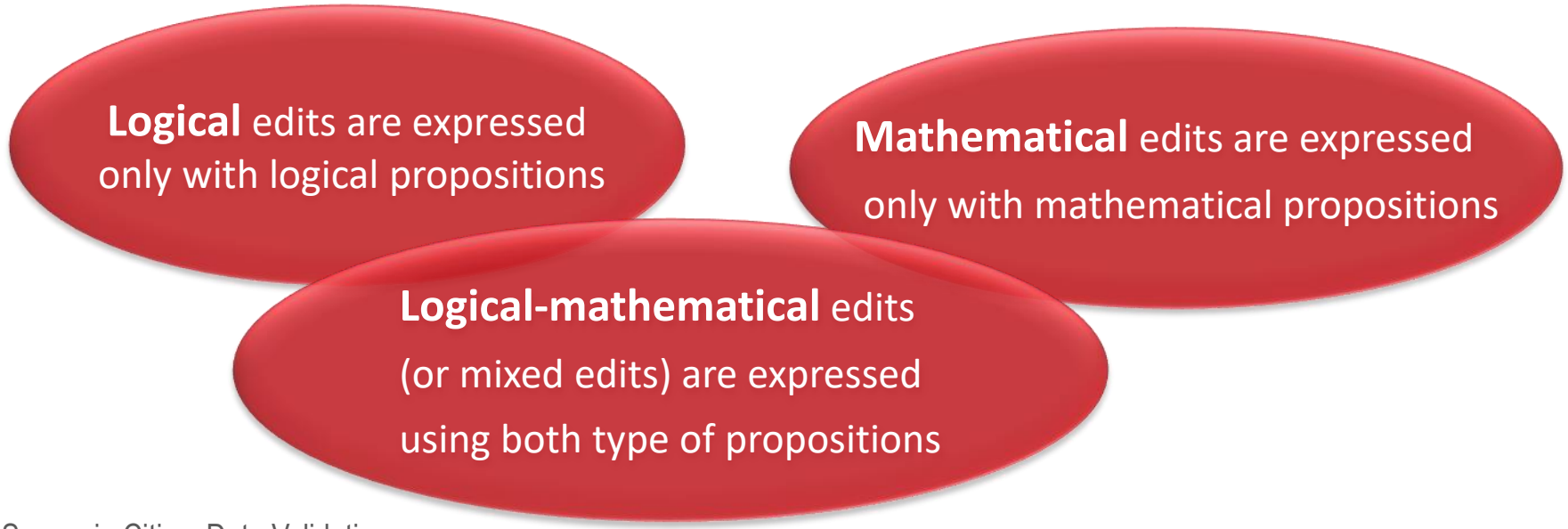
  Examples:

  $(x_1 < 14)$

- *Mathematical proposition* expresses a mathematical condition on values of at least two quantitative fields

  Example:

  $(x_1 - x_2 \geq 14)$



A generalized data editing for error detection (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 PM10 number of exceedances of the daily average of 50 µg/m3 is ≥ 0, then PM10 annual average concentration value should be ≥ 0

can be written as

$\overline{A}$ OR B

A ⇨ B

Consistency edit

- PM10_SUP_CENTR_ARIA<0 OR PM10_MEDIA_CENTR_ARIA_T1≥0

Conflict edit

- PM10_SUP_CENTR_ARIA≥0 AND PM10_MEDIA_CENTR_ARIA_T1<0

A AND $\overline{B}$

Conclusions

The main objective of the project will be:

- Design and implement the validation process from a generalized perspective, including reporting for error evaluation and analysis.
- Define the methodologies and algorithms necessary to carry out the automatic checks required by the validation process.
- Design and develop application components and database, including procedures for integrating the validation process with the capture and production environment.
- Analyze and validate the tools developed and the results of the new validation process through the definition of test cases and ongoing experimentation on the 2023 survey.

Thank you for your attention

Domenico Adamo adamo@istat.it

Gianpiero Bianchi gianbia@istat.it

Paolo Francescangeli pafrance@istat.it

Lucia Mongelli mongelli@istat.it

Istat | Istituto Nazionale di Statistica