INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

eurostat

The conference is partly financed by the European Union

EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS

2024 ESTORIL - PORTUGAL

A step-by-step process to deal with the protection of a set of tabular data

Julien Jamme, Insee

Clara Baudry, Insee

Insee

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

eurostat

The conference is partly financed by the European Union

Context

- Very large tabular data dissemination
- SDC with suppressive methods
- High sensitivity of the protection step to the actual dissemination content
- Protection methodology and tools require high level of expertise

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL
eurostat
The conference is partly financed by the European Union

Issues

- Problems not generalized
- Processes not automated
- Complexity of the task

=> Protection process supported by a team in the department of statistiscal methods

The conference is partly financed by the European Union

Objective

- The long-term goal : Let the producers protect data on their own
- To reach it :

  - Develop methodology and tools to reduce the expertise level required to handle protection of large tabular data dissemination

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL
eurostat
The conference is partly financed by the European Union

Methodological approach

- ***Describe*** the process step by step to identify issues and sticking points
- ***Generalize*** each step to handle every case of dissemination
- ***Automate*** where possible to reduce the implementation burden and make the code reproducible as much as possible
- ***Disseminate*** methods, tools and practices by training people

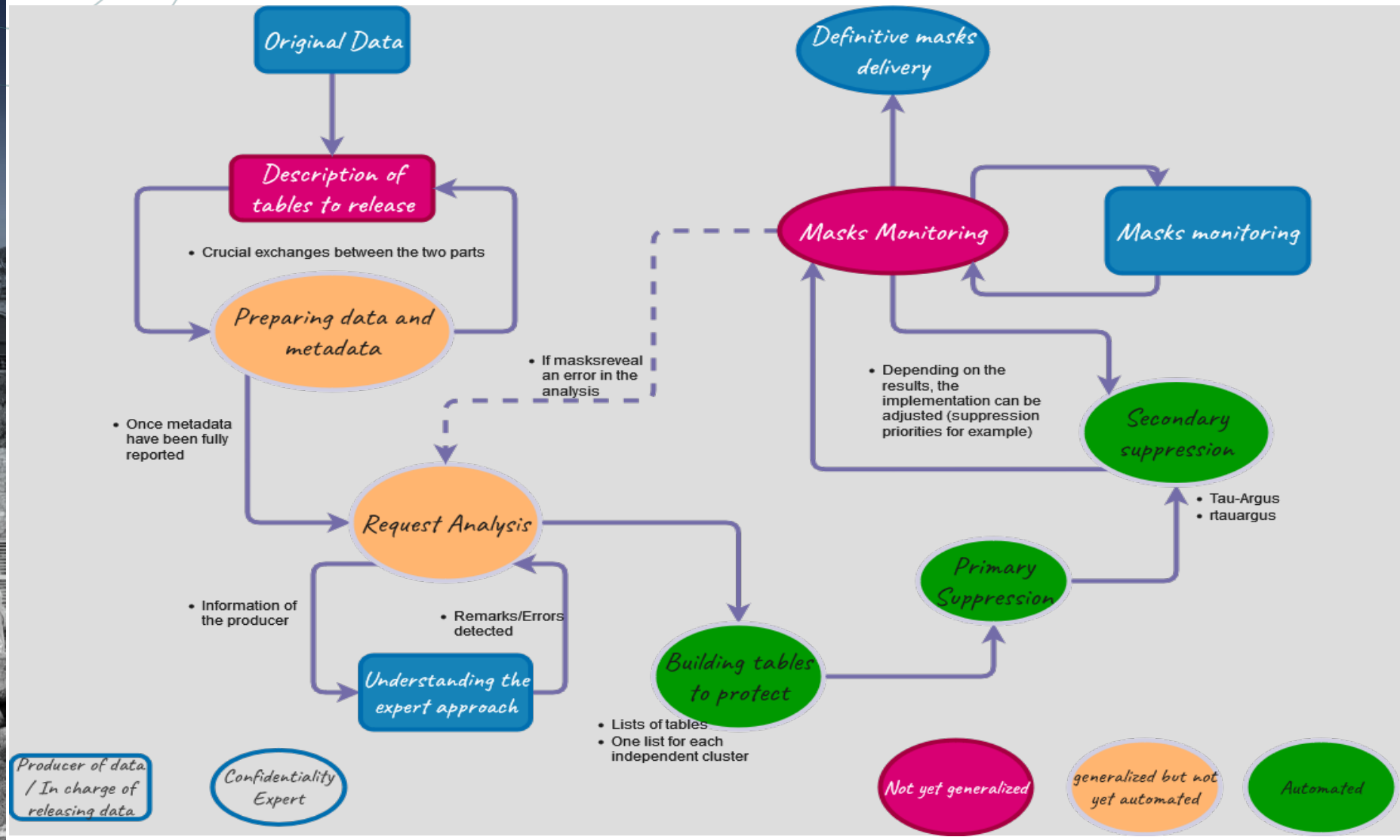
Methodology improvements while producing confidentiality masks

Ongoing work

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL
eurostat
The conference is partly financed by the European Union

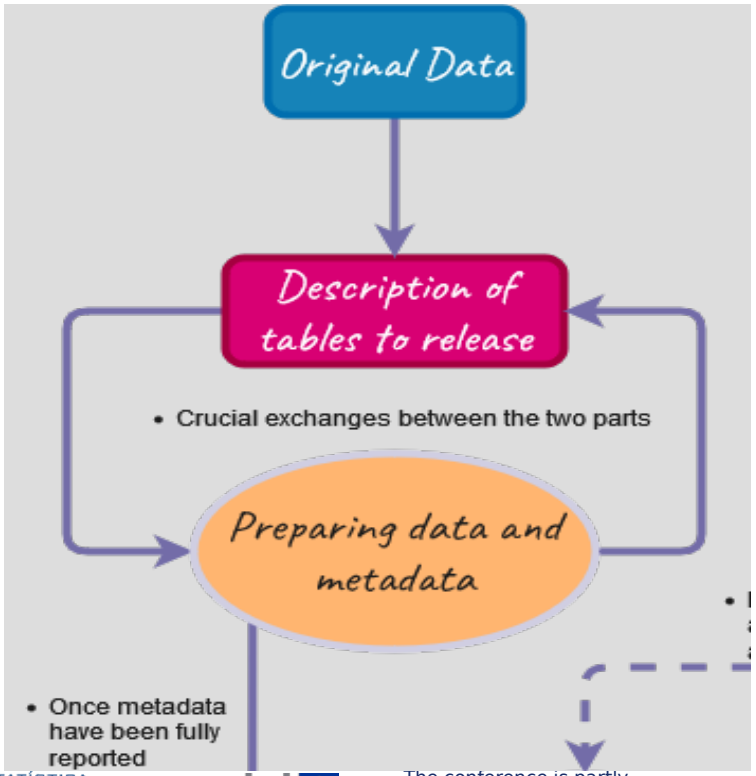

EUROPEAN CONFERENCE ON
QUALITY IN OFFICIAL STATISTICS
2024 ESTORIL - PORTUGAL

Describe


Original Data
Description of tables to release
• Crucial exchanges between the two parts
Preparing data and metadata
• Once metadata have been fully reported
Request Analysis
• Information of the producer
Understanding the expert approach
• Remarks/Errors detected
• If masksreveal an error in the analysis
Definitive masks delivery
Masks Monitoring
Masks monitoring
• Depending on the results, the implementation can be adjusted (suppression priorities for example)
Secondary suppression
• Tau-Argus
• rtauargus
Primary Suppression
Building tables to protect
• Lists of tables
• One list for each independent cluster
Producer of data / In charge of releasing data
Confidentiality Expert
Not yet generalized
generalized but not yet automated
Automated

Issues and Sticking points

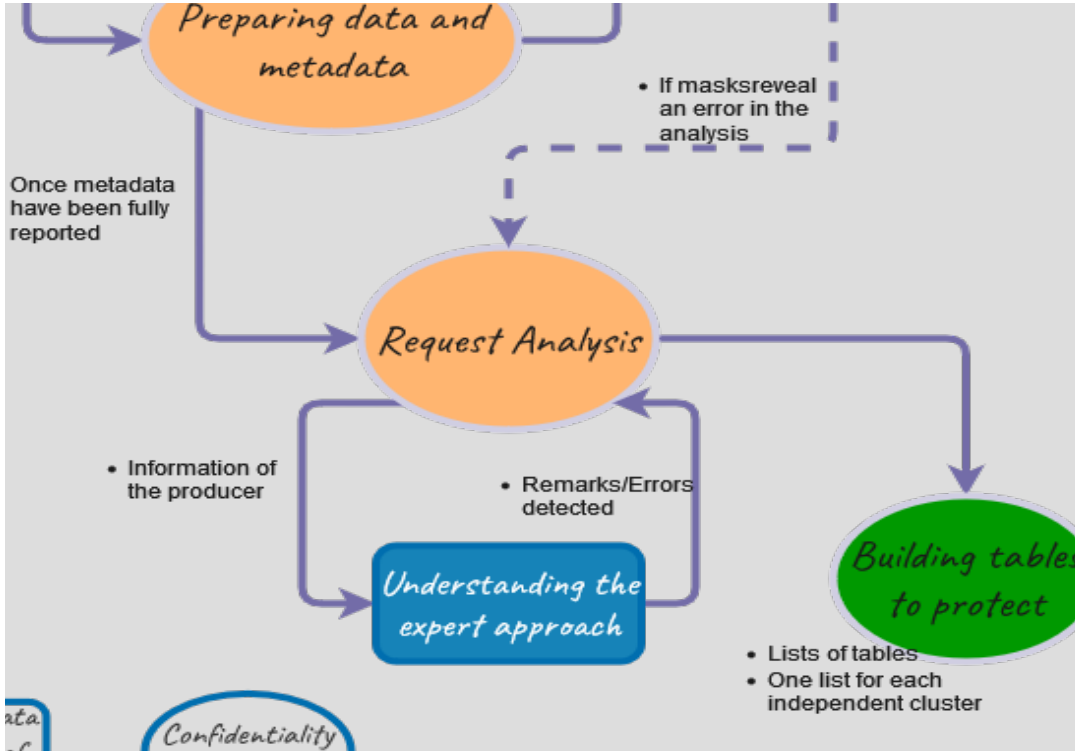
1) The initial contact between producer and confidentiality expert



INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL
eurostat
The conference is partly financed by the European Union

Issues and Sticking points

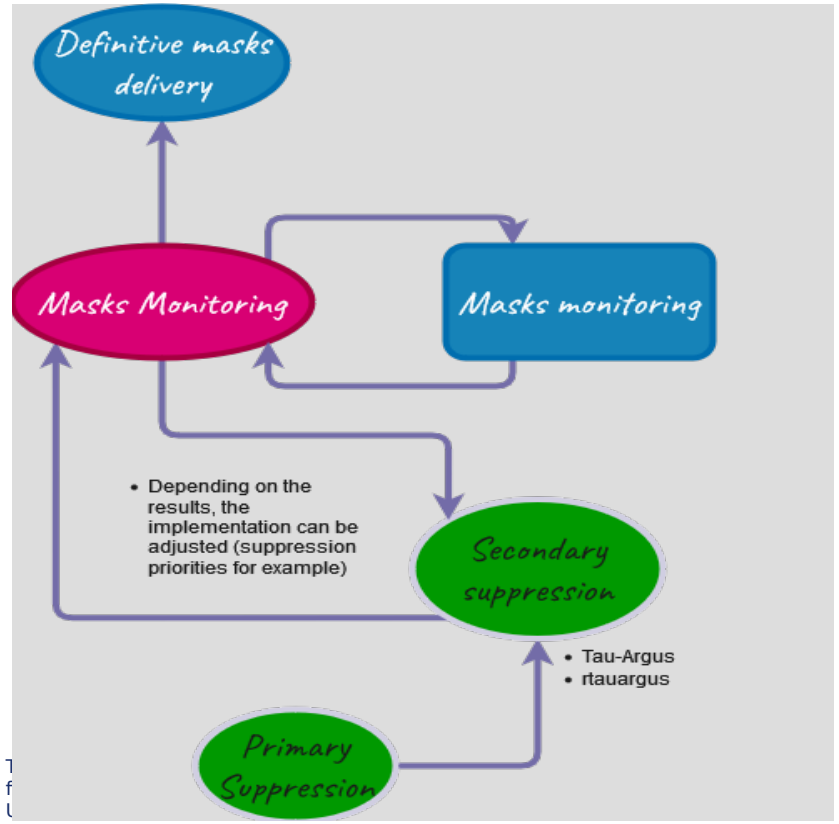
2) The analysis of the dissemination



INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL
eurostat
The conference is partly financed by the European Union

Issues and Sticking points

3) The implementation of the suppression process on very large dissemination



Step 1 : Building relevant metadata

- First issue : Make the initial contact productive
- Objective :

  - Producer : to give a complete description of the dissemination

  - Expert : to get all the information necessary for the protection process (and no more)

- data expert vs confidentiality expert

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL
eurostat
The conference is partly financed by the European Union

Step 1 : Building relevant metadata

- Find a way to universally describe a table (from a protection perspective)
- What is required ?

Indicator or Response Variable

Number of companies ***by region and CEO's gender in the industrial sector***

Crossing variable n°1

		Headquarter's region				
		North	West	East	South	**Total**
Gender of the CEO	Female	6	14	1	7	**28**
	Male	3	2	1	13	**19**
	Total	**9**	**16**	**2**	**20**	**47**

Cell

Field: *Industrial sector Companies* ***– Source : Antipol 2052***

Crossing variable n°2

Field

Step 1 : Building relevant metadata

- Find a way to universally describe a table (from a protection perspective)
- Formalize the description

$$\mathrm{RV}^{hrc_{RV}} \underset{Field^{hrc_F}}{\otimes} \left\{ \mathrm{CV1}_{tot_1}^{hrc_1} \times \mathrm{CV2}_{tot_2}^{hrc_2} \right\}$$

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL
eurostat
The conference is partly financed by the European Union

Step 1 : Building relevant metadata

- Find a way to universally describe a table (for the protection perspective)
- Formalize the description

$$\mathrm{Freq} \underset{Ind.Sect}{\otimes} \{\mathrm{Region}_{total} \times \mathrm{Gender}_{total}\}$$

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL
eurostat
The conference is partly financed by the European Union

Step 1 : Building relevant metadata

- Relevant metadata contain a complete description of all tables :
  - spanning and response variables
  - field definition
  - hierarchies (nested or non-nested),
  - etc.
- They must allow to analyze the links between the tables

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL
eurostat
The conference is partly financed by the European Union

Step 1 : Building relevant metadata

- Producer has to be able to fill in the file
  => Choice for a spreadsheet format

Table	*Field*	*Response Variable (RV)*	*Crossing Variable (CV)*	*CV Total Code*
T1	Companies of Industrial Sector	Frequencies	HQ's Region	Total
T1	Companies of Industrial Sector	Frequencies	CEO's Gender	Total

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL
eurostat
The conference is partly financed by the European Union

Step 2 : Analyzing the dissemination

- Objective of the analysis :
  - Determine the links between tables
  - Linked tables => simultaneaous protection
  - Unlinked tables => independent protection
  - Ultimate goal : describe the tables on which suppression is really applied

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL
eurostat
The conference is partly financed by the European Union

Step 2 : Analyzing the dissemination

- Issues
  - High expertise and experience required
  - Very sensitive to the actual content
  - Lack of a methodology to handle this step
  - Fit the method to every situation (non nested hierachies, holding variable, links on response or spanning variable, complementary fields, etc.)

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL
eurostat
The conference is partly financed by the European Union

Step 2 : Analyzing the dissemination

- Our proposition
  Build a complete methodology to analyze dissemination :
  - By describing steps and objectives
  - By generalizing the approach to handle every situation as much as possible
- Automation of the analysis is under way

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL
eurostat
The conference is partly financed by the European Union

Step 2 : Analyzing the dissemination

$$\begin{cases} \text{to\_margarita} \otimes \{\text{NUTS2} \times \text{SIZE}\} \\ \text{to\_margarita} \otimes \{\text{NUTS3} \times \text{SIZE}\} \\ \text{to\_calzone} \otimes \{\text{NUTS2} \times \text{SIZE}\} \\ \text{to\_calzone} \otimes \{\text{NUTS3} \times \text{SIZE}\} \\ \text{to\_pizzas} \otimes \{\text{NUTS2} \times \text{SIZE}\} \\ \text{to\_pizzas} \otimes \{\text{NUTS3} \times \text{SIZE}\} \end{cases}$$

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

eurostat

The conference is partly financed by the European Union

Step 2 : Analyzing the dissemination

$$
\begin{cases}
\text{to\_margarita}^{pizzas} \otimes \left\{ \text{NUTS2}_{all}^{nuts} \times \text{SIZE}_{all} \right\} \\
\text{to\_margarita}^{pizzas} \otimes \left\{ \text{NUTS3}_{all}^{nuts} \times \text{SIZE}_{all} \right\} \\
\text{to\_calzone}^{pizzas} \otimes \left\{ \text{NUTS2}_{all}^{nuts} \times \text{SIZE}_{all} \right\} \\
\text{to\_calzone}^{pizzas} \otimes \left\{ \text{NUTS3}_{all}^{nuts} \times \text{SIZE}_{all} \right\} \\
\text{to\_pizzas}^{pizzas} \otimes \left\{ \text{NUTS2}_{all}^{nuts} \times \text{SIZE}_{all} \right\} \\
\text{to\_pizzas}^{pizzas} \otimes \left\{ \text{NUTS3}_{all}^{nuts} \times \text{SIZE}_{all} \right\}
\end{cases}
$$

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL
eurostat
The conference is partly financed by the European Union

Step 2 : Analyzing the dissemination

$$\begin{cases} \text{to\_margarita}^{pizzas} \otimes \{\text{NUTS}_{all}^{nuts} \times \text{SIZE}_{all}\} \\ \text{to\_calzone}^{pizzas} \otimes \{\text{NUTS}_{all}^{nuts} \times \text{SIZE}_{all}\} \\ \text{to\_pizzas}^{pizzas} \otimes \{\text{NUTS}_{all}^{nuts} \times \text{SIZE}_{all}\} \end{cases}$$

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL
eurostat
The conference is partly financed by the European Union

Step 2 : Analyzing the dissemination

$$\text{to} \otimes \left\{ \text{NUTS}_{all}^{nuts} \times \text{SIZE}_{all} \times \text{PIZZAS}_{pizzas}^{(h)} \right\}$$

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL
eurostat
The conference is partly financed by the European Union

Step 3 : Suppression

- Initially
  - the most intensive step in terms of code
  - One dissemination => One algorithm to implement to handle all the links between tables
- Context
  - Tau-Argus as reference
  - But some limitations with numerous linked tables dissemination

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL
eurostat
The conference is partly financed by the European Union

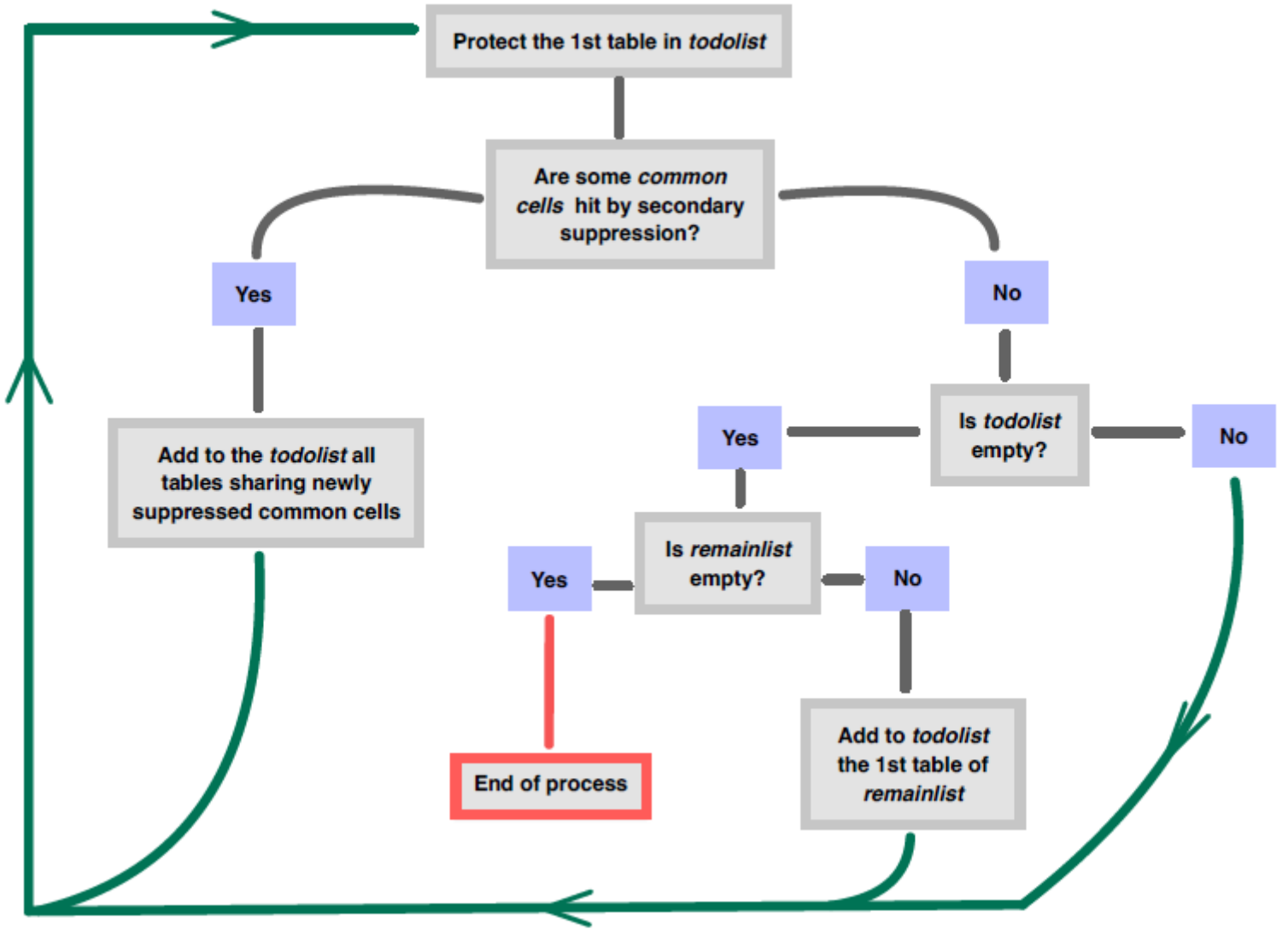
Step 3 : Suppression

- Solution
  - **rtauargus** : R package developed to interface R with Tau-Argus
  - + algorithm implemented to manage the linked tables
  - Versionning with **git**

- Productivity and efficiency gain
- Reduction of code length

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL
eurostat
The conference is partly financed by the European Union

Step 3 : Suppression



Review

- Description of the whole process => to identify issues to be addressed
- Generalization :
  - Formalization of table description
  - Building of metadata file
  - Describing steps to analyze a dissemination
- Automation :
  - Analysis step (under way)
  - Suppression step (completed)
- Dissemination of expertise :
  - Internal training
  - Guidelines

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL
eurostat
The conference is partly financed by the European Union

Outlook

How far are we from the ultimate goal of our quest ?

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL
eurostat
The conference is partly financed by the European Union

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

eurostat

The conference is partly financed by the European Union

EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS

2024 ESTORIL - PORTUGAL