Istat | Istituto Nazionale di Statistica

Rome, 1-2 December 2025

4th Workshop on
Methodologies For Official Statistics

Methodological Framework for developing Open Source Software Tools at Istat

Marcello D'Orazio

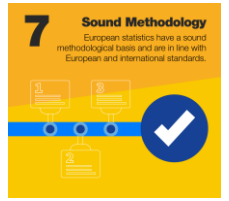
Italian National Institute of Statistics - Istat | Directorate for Methodology and Statistical Process Design

Why Open Source Statistical Tools (1/2)

- Nowadays producing official statistics in NSI requires highly complex processes due to the need of integrating different data sources (administrative registers, survey data, big data).

Specify needs	Design	Build	Collect	Process	Analyse	Disseminate	Evaluate

- This calls for
  - efficient and **cost-effective statistical IT tools** (principle 10 of the EU CoP)
  - that ensure the implementation of **sound methodologies** (principle 7 of the EU CoP)



- Istat has used and still uses **SAS** software for various tasks.
- To face challenges of modernization of processes, about 20 years ago Istat started to explore statistical Open Source Software Tools (OSST)

The choice fell on the R environment (R Core Team, 2025) https://www.R-project.org/

Istat

Why Open Source Statistical Tools (2/2)

- Decision to adopt OSST:

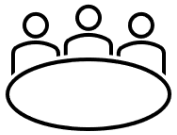
also in response to a ministerial directive (April **2003**) inviting government agencies to adopt OSST and avoid dependence on a single commercial software tool

Recently reinforced by the **2019** guidelines issued by the Agency for Digital Italy (AgID)

- **EU** also promoted and is promoting the adoption of OSST in its "Open Source Software Strategy" (rel. 2014 for period 2014-2017, updated in 2020 for period 2020-2023).

It proposes a "Think Open" approach aimed at promoting the sharing and reuse of OSST

Open Source Statistical Tools: Work Done at Istat (1/2)

- Introduction of R took place in the methodology directorate shortly after the year **2000**, mainly to test new methods
- An informal group of R experts was created 
- Regular training for staff started in **2007** (now "base R" and "intermediate R" and various courses on specific topics/packages). In addition, R tools are used in statistical courses (sampling, data integration, data processing and imputation, etc.)
- R was chosen to create new packages or to migrate old stand-alone tools or tools developed in other languages. The first R packages were released in **2010**
- Introduction of Python (around **2020**) to apply Machine Learning algorithms & use of "non-traditional" data sources

Open Source Statistical Tools: Work Done at Istat (2/2)

Some of the developed R packages

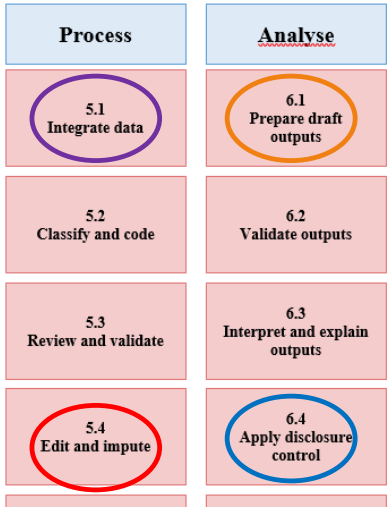
GSBPM	R package	Description
2.4 Design frame and sample	**R2BEAT**	Multistage Sampling Allocation and Sample Selection
	SamplingStrata	Optimal Stratification of Sampling Frames for Multipurpose Sampling Surveys
5.1 Integrate data	**StatMatch**	Methods for Statistical Matching (aka Data Fusion) and hotdeck imputation of missing values
5.4 Edit and impute	**SeleMix**	Selective Editing via Mixture Models. Detection of outliers and influential errors using a latent variable model.
5.6 Calculate weights & 5.7 Calculate aggregates	**ReGenesees**	R evolved Generalised software for sampling estimates and errors in surveys (calibration, estimation and sampling error assessment)

Other tools (R and others) can be found in the Istat repository on the corporate website:

https://www.istat.it/en/classifications-and-tools/methods-and-software-of-the-statistical-process/

Open Source Statistical Tools: Ongoing Work at Istat

- **Support** the introduction of statistical OSST into regular ongoing statistical production processes
- **Migration** of obsolete tools in open source tools or adoption of OSST developed externally:



  - GSBPM 5.1 **Integrate data**: new tool for record linkage based on existing R packages
  - GSBPM 5.4 **Edit and Impute**
    - Test R packages **validate**, **validatetools**, **errorlocate**, and **VIM** → **ARTIC**
    - New R package **UnitMix**
  - GSBPM 6.1. **Prepare draft outputs**: seasonal adjustment → extending the functionalities of the **RJDemetra**
  - GSBPM 6.4 **Apply disclosure Control**: test R packages **ptable**, **cellKey**, **rtauargus** etc.

Open Source Statistical Tools: Governance

Our experience shows that in official statistics the advantages of statistical OSST outweigh the disadvantages

However, this argument alone cannot be used to promote their adoption in the statistical production processes; it is necessary to consider the technical, organisational and cultural levels.

It is necessary to establish a strategy that should include:

- procedures for assessment of the potential tools (including code security)
- guidelines for code development and its documentation
- testing procedures
- training programs for staff
- communication initiatives to share best practices and successful use cases
- establishment of internal expertise and support structures
- cooperation and partnership with other NSIs or EU statistical offices on key projects/tests or to share knowledge and expertise
- engagement with open source communities
- development of corporate policies and governance bodies

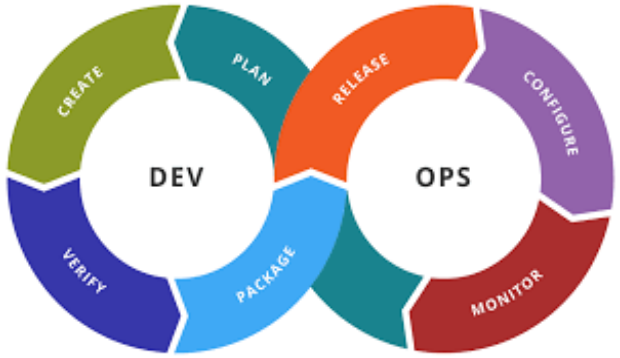
Open Source Statistical Tools: Governance

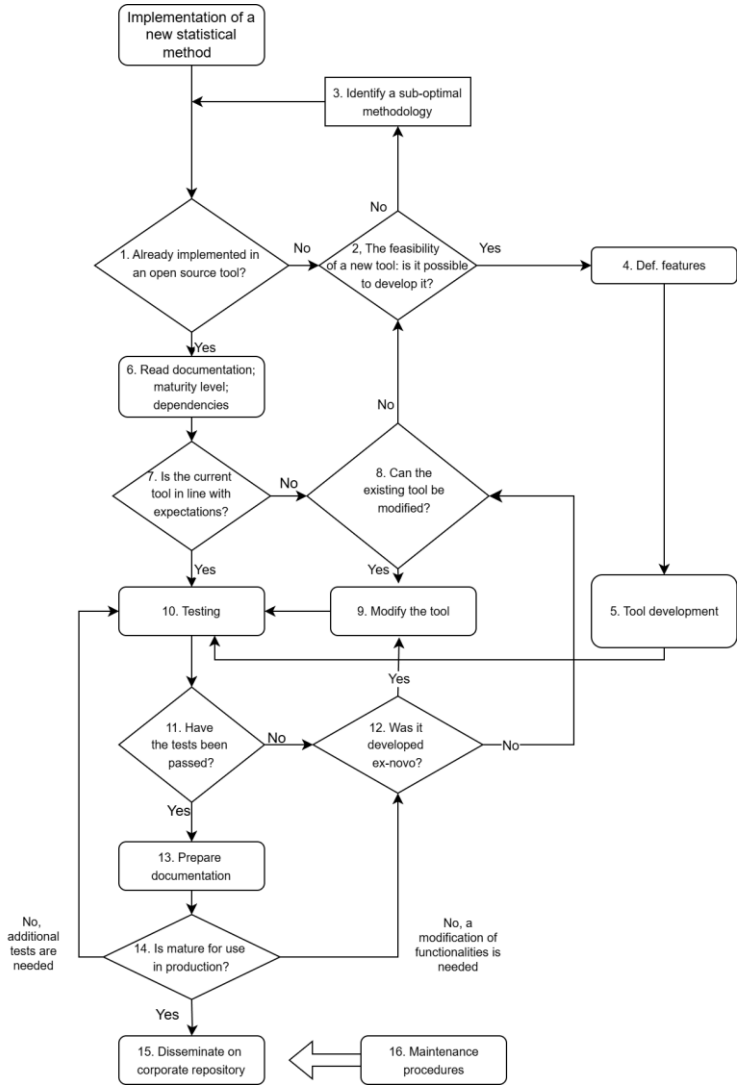
- procedures for assessment of the potential tools (including code security);
- guidelines for code development and its documentation;
- testing procedures; 😐

Two docs released in Dec 2024:
- ✓ Guide to R code devel. e doc.
- ✓ Corporate methodological governance of statistical OSST

- training programs for staff ☺
- communication initiatives to share best practices and successful use cases ☺ (internal webinars)
- establishment of internal expertise and support structures ☺ (Istat's network of R experts)
- cooperation and partnership with other NSIs or EU statistical offices on key projects/tests or to share knowledge and expertise ☺
  - EU EG on "Shared Tools" (from end of 2023; continuation of OS4OS of 2021-22);
  - UNECE EG on Open Source Software in 2024 https://unece.github.io/OSS/
- engagement with open source communities 😐
  - "regular" participation at **Use of R in Official Statistics – uRos** conference https://r-project.ro/conferences.html
- development of corporate policies and governance bodies 😐 (some policies developed, govern. bodies to be created)

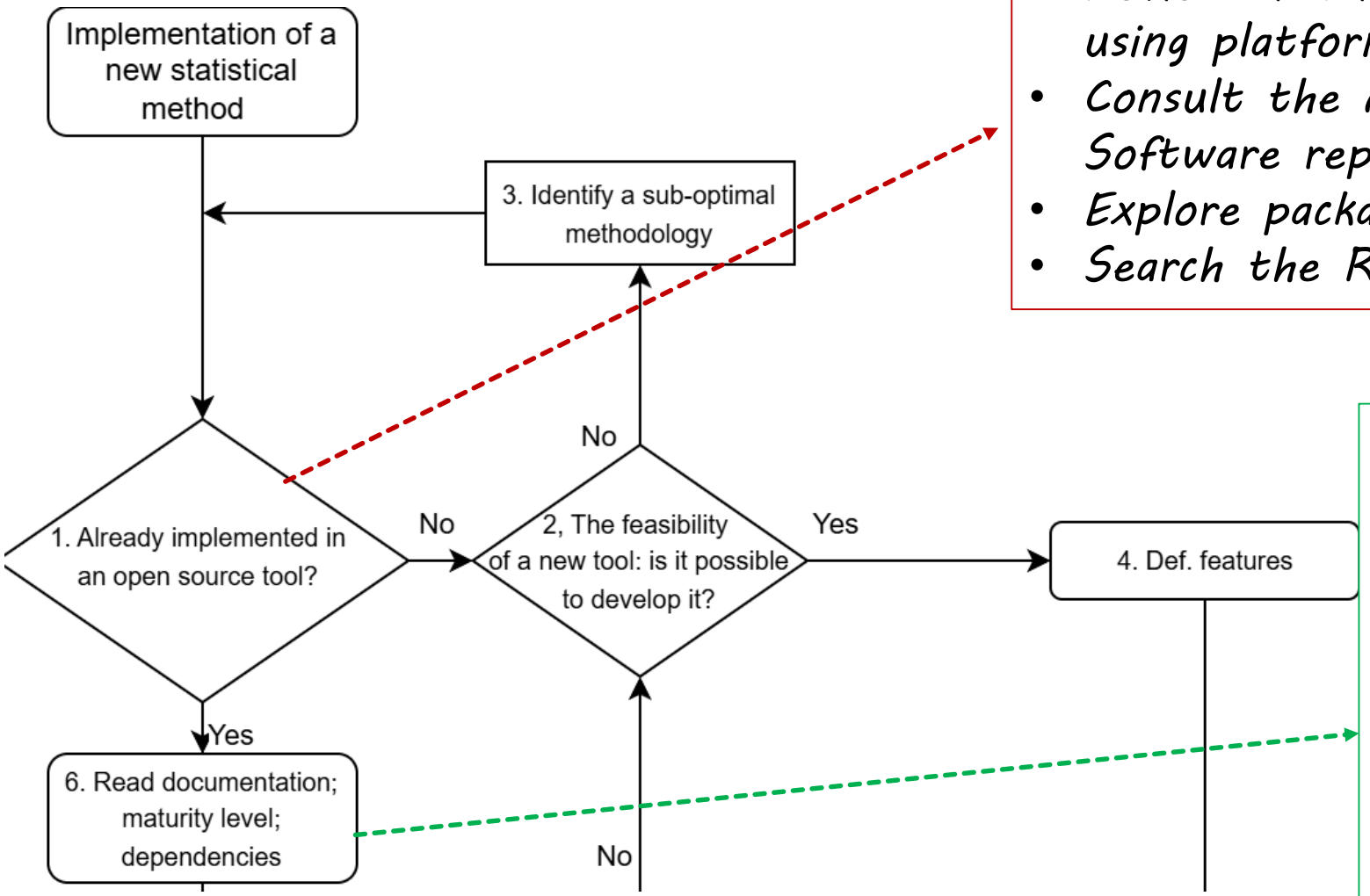
Governance of statistical OSST: methodological governance





D'Orazio et al (2025) Statistical Journal of the IAOS

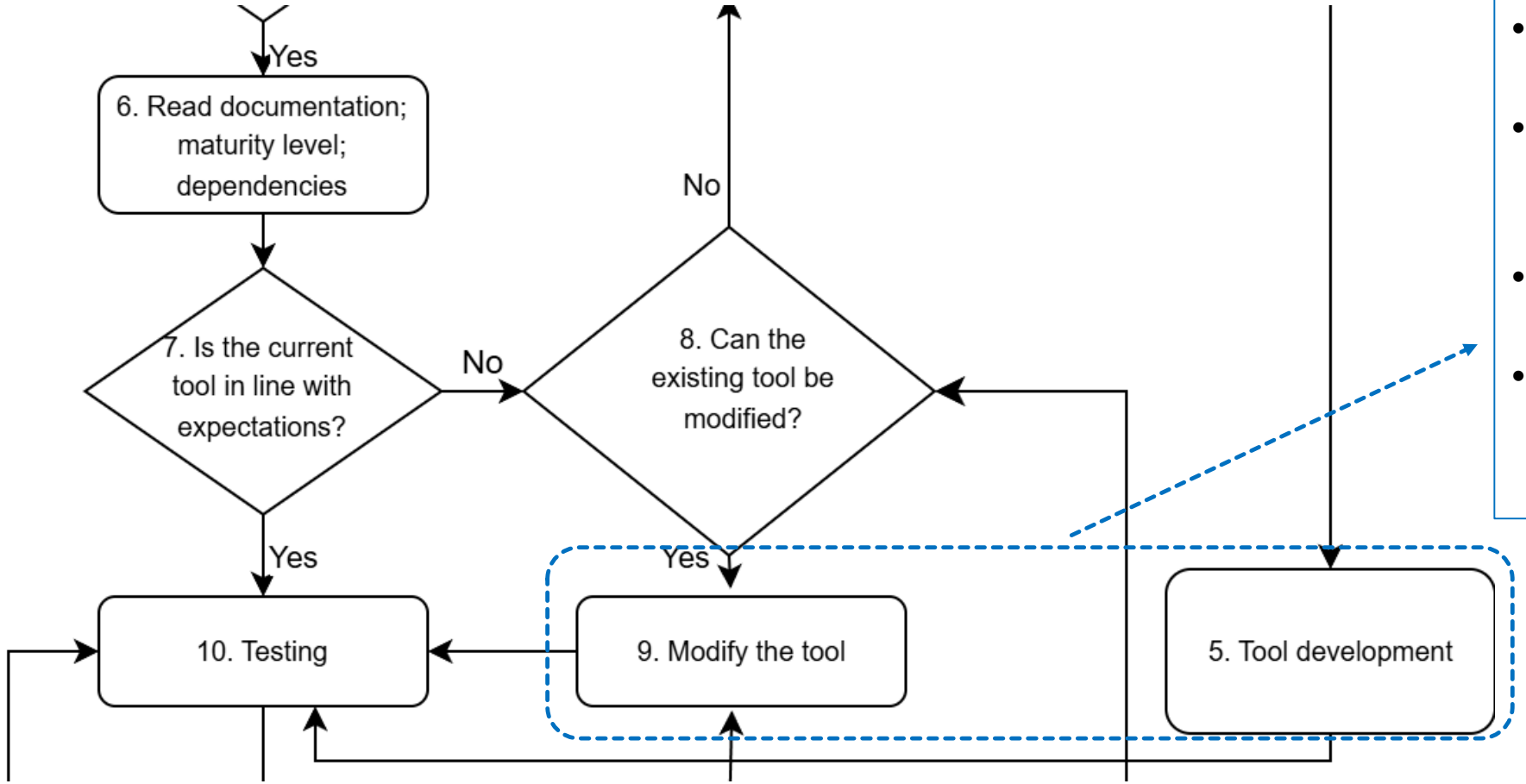
Governance of statistical OSST: methodological governance



- Review citations in methodological papers using platforms like Google Scholar.
- Consult the Awesome Official Statistics Software repository
- Explore packages available on CRAN
- Search the R website using Google

- Prefer tools developed by EU institutions, NSIs, etc
- Evaluate the tool's maturity
- Review the history; bug fixes; reasons of changes → Prefer most stable tool
- Avoid tools with excessive dependencies

Governance of statistical OSST: methodological governance

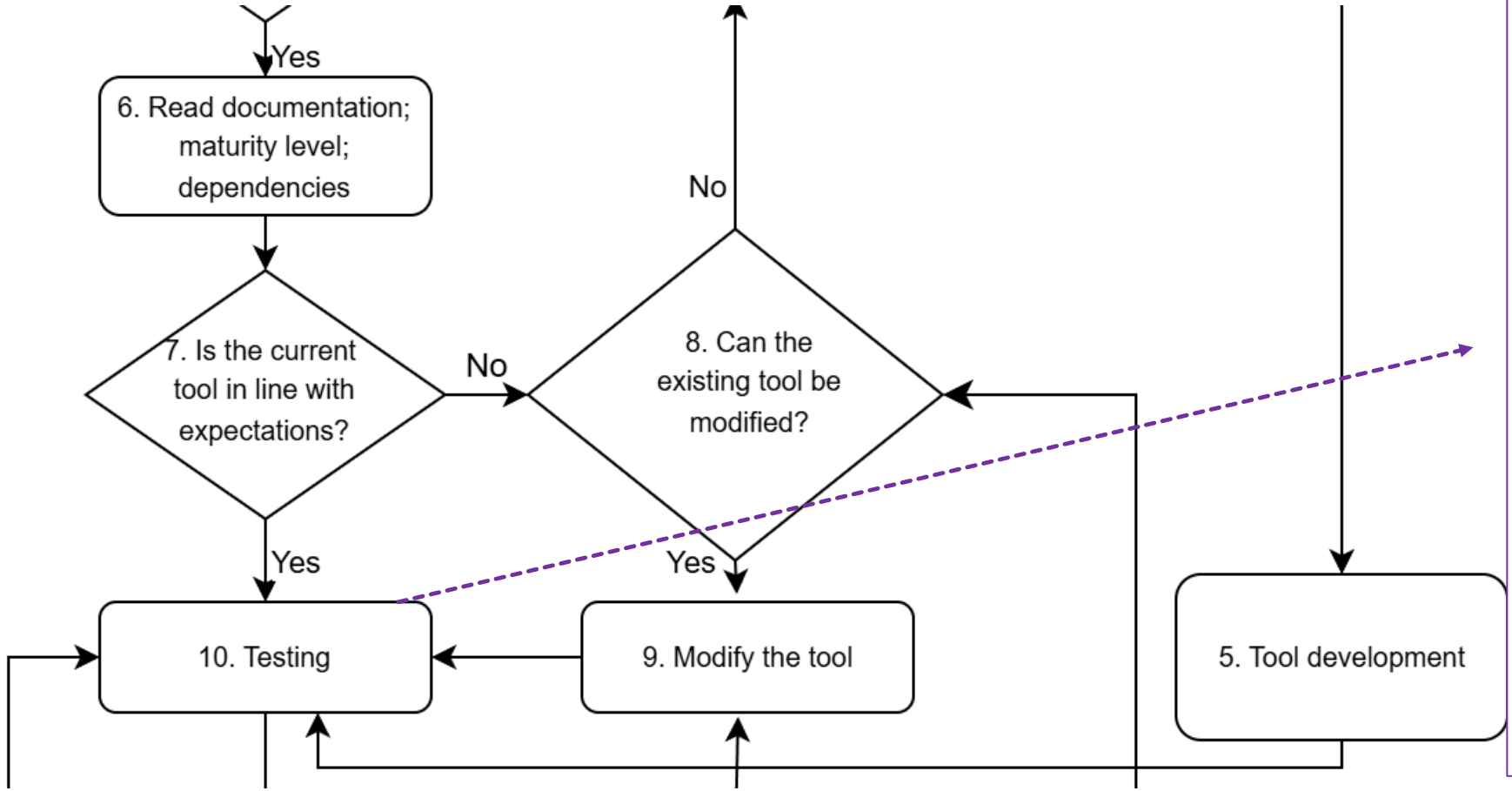


For R

- Rules for writing and commenting code
- Documenting new functions → **roxygen** Skeleton
- Use of version control (git – GitLab)
- Writing extended documentation with R Markdown or quarto

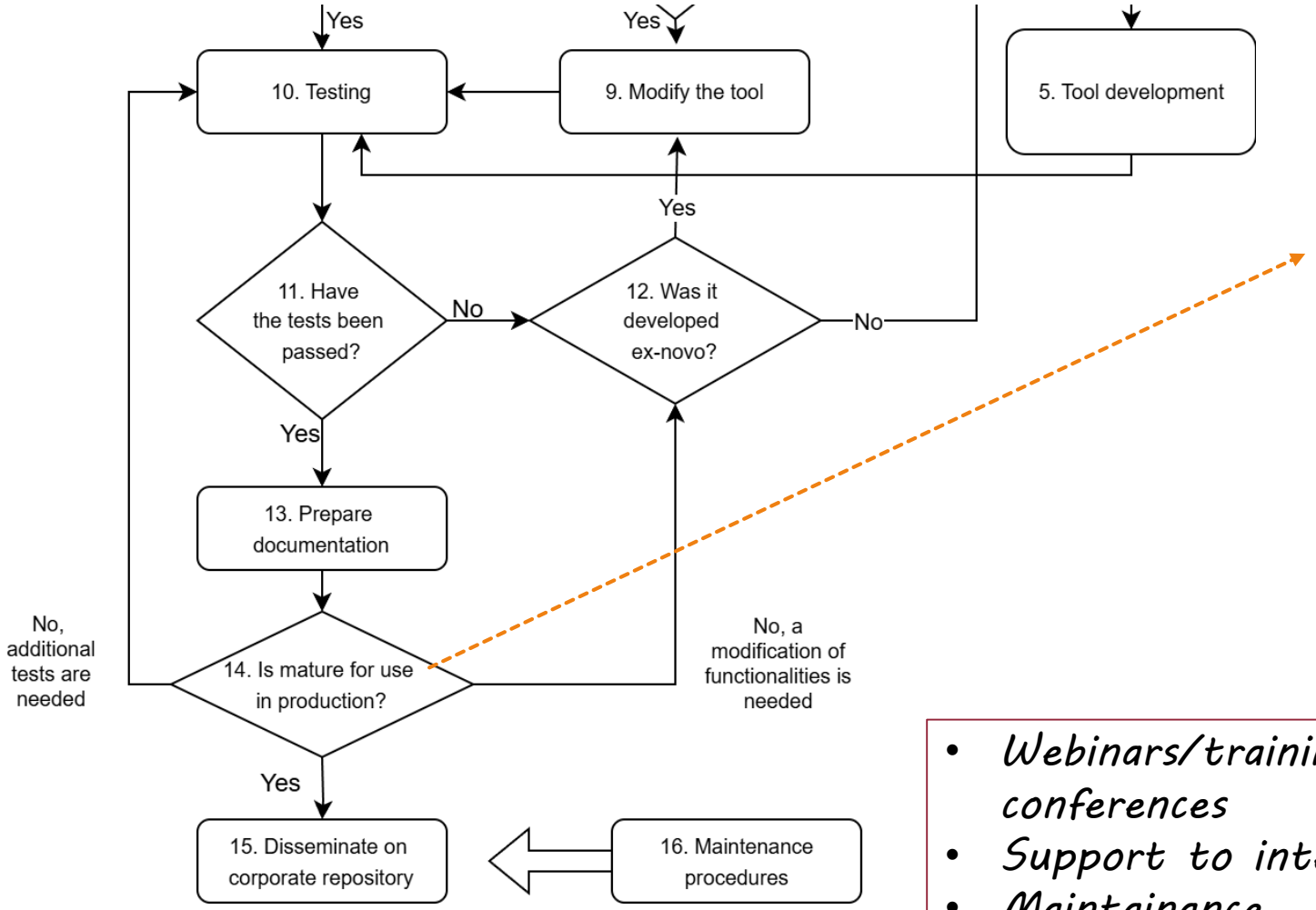
Istat

Governance of statistical OSST: methodological governance



- *From simple to increasingly complex scenarios*
- *Start with trials on artificial or real datasets (results are easy to verify)*
- *Then test on subsets of real data*
- *Go in a full production cycle; assess accuracy and computational perform.*
- *If the tool replaces another one, conduct parallel testing to compare results*

Governance of statistical OSST: methodological governance



- *The tools is presented in dedicated meetings, covering features, development process, and test results*
- *A committee with experts in statistical methodology and software development·evaluates the tool*
- *The committee decides whether to approve the tool for use in production or requests further testing or changes·*

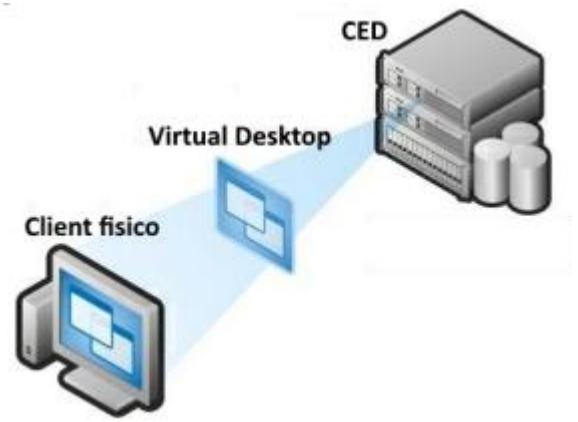
- *Webinars/trainings for staff; presentation at conferences*
- *Support to internal and external users*
- *Maintainance*
- *Ensure continuity of maintenance*

Governance of statistical OSST: IT governance

To be started!

Major challenges

- optimize use of R and associated tools (Rstudio/Positron etc.) in a Virtual Desktop Interface (VDI comes with a cloud-based online storage and file hosting service)
- interoperability (Oracle, SAS, R, Python, …)



Task	RSUBMIT Used?	SAS/IML Needed?	Method
Run SAS code remotely	✔	✘	RSUBMIT
Run R locally from SAS	✘	✔	SAS/IML + `submit / R`
Run R remotely	✔	✔	RSUBMIT + SAS/IML
Run Python locally	✘	✘	`proc python`
Run Python remotely	✔	✘	RSUBMIT + `proc python`

MicroServices

- **small, independent services**, each responsible for a specific function
- Services can use **different programming languages** or databases
- services communicate with each other through **lightweight, standardized interfaces**, usually APIs

Thank You for Your Attention!

Marcello D’Orazio | marcello.dorazio [at] istat.it

Istat | Istituto Nazionale di Statistica