

UNITED NATIONS  
ECONOMIC AND SOCIAL COMMISSION FOR ASIA AND THE PACIFIC (ESCAP)  
STATISTICAL INSTITUTE FOR ASIA AND THE PACIFIC (SIAP)  
In partnership with  
THE VANUATU BUREAU of STATISTICS (VBoS)

**e-Learning Course**

**Principles of Reproducible Analytical Pipelines for Official Statistics**

**6 May 2024 – 31 May 2024**

**I. About the Course**

This e-learning course aims to build capacity in national statistical systems for the development and implementation of Reproducible Analytical Pipelines (RAPs) for Official Statistics.

*What is a Reproducible Analytical Pipeline?*

Simply put, reproducible analytical pipelines (RAPs) are automated statistical processes (data processing and analysis) that codify to the greatest extent possible the production of official statistics. Common tools that are used to develop RAP include software such as R or Python, and version control management tools such as Git.

Reproducibility is at the heart of the approach. It implies that the outputs can be generated again with any new or revised input datasets using the RAP developed. This also implies drafting documents explaining the RAP that make it possible to build institutional knowledge and use the RAP in the future by new staff.

*Why are Reproducible Analytical Pipelines important for Official Statistics?*

All national statistical systems are engaged in the regular, high frequency production of many official statistics. For example, most countries compile monthly consumer price index (CPI). The input data for the compilation of CPI is generally the same from month to month. By developing and implementing an RAP for the compilation of CPI, countries can improve the timeliness and quality of the CPI since automation reduces the time required to clean and analyze the data; it also reduces the chance of errors that could occur when relying on non-automated processes.

Furthermore, the Sustainable Development Goals (SDGs) require that countries use more diverse data sources in the compilation of indicators. The COVID-19 crisis has shown that automated tools can facilitate data analysis and reporting when these sources are updated. These tools, including software such as R and sharing platforms such as GitHub, allow statisticians to streamline data cleaning, compilation, and analysis.

Implementing these new approaches to data processing and analysis and utilizing these modern tools can aid in the timelier production of statistics and lower the time staff need to produce statistics. This approach is also more transparent, easier to review and to share among staff and thus less prone to errors in the data and/or statistical analysis. Furthermore, implementing good practices based on RAP can contribute to building institutional knowledge by ensuring that work programmes can be more seamlessly transferred among different staff and over time.

## II. Target Audience

The main target audience for the course is technical staff working in statistical offices that serve in roles of data processors or data analysts. Since RAP is a process that involves data from various sources, data producers and analysts from the broader national statistical system who are interested in RAP are welcome to enroll in the course.

It is important to note that since this course has a practical goal, participants should have some experience in R programming. **[A mandatory tutorial](#) is therefore required for participants to complete before being able to access the 4 modules of the course.**

**The workload of this course is 3h/week.** It is thus important for participants to organize themselves and allocate enough learning time during the course.

## III. Learning Objectives

At the end of the course participants should be able to:

- Identify and assemble the critical elements for creating a RAP
- Appreciate the logic of the RAP process as a whole
- Master the principles of Reproducible Documents
- Use R and R Markdown to produce an automatic report
- Describe the steps involved in the production of a periodic report.
- Recognize the main principles and actions of Git And GitHub for file versioning
- Adopt good practices and lessons learned by early adopters of RAP when developing RAP
- 

## IV. Course Design and Content

Since R will be used throughout this course, participants should complete a tutorial that will be accessible before the start of the course. This tutorial provides guidance and resources on R, how to install it , and how to start with RStudio. The tutorial also shows, through interactive exercises, the main logic of R programming and provides some simple coding lessons as a refresher in R.

The course will include 5modules including the R tutorial.

Module	Coverage
M0: R and RStudio prerequisites	<ul style="list-style-type: none"> <li>• How to install R and RStudio</li> <li>• Install packages</li> <li>• Run a few line of code</li> </ul>
M1-Why NSOs should invest in RAP?	<ul style="list-style-type: none"> <li>• What is a RAP?</li> <li>• Why Reproducibility is important</li> <li>• What are some of the tools that we will learn more about in this course?</li> <li>• Case Study: The example of Vanuatu RAP</li> </ul>
M2-Tools for RAP	<ul style="list-style-type: none"> <li>• From Data to R to R markdown: Knitting text and code together</li> <li>• Basics of R markdown and automatic reports</li> <li>• Introduction to Version control</li> <li>• Case Study: The technical workflow of Vanuatu RAP</li> </ul>
M3-Reproducible documents	<ul style="list-style-type: none"> <li>• Principles and good practices for reproducibility</li> <li>• File management &amp; naming rules</li> <li>• Functions in R</li> <li>• Creation of a fully automated report</li> </ul>
M4-Putting it All Together	<ul style="list-style-type: none"> <li>• Git and GitHub for RAP</li> <li>• Working with version control in RStudio</li> <li>• Case Study: A RAP for working with Trade data at VBoS</li> <li>• To go further: Controlling environments, advanced tools &amp; methods.</li> </ul>

## V. Evaluation

The evaluation will be based on the learner's ability to apply the methods and training materials described during the course, as well as on their capacity to recall the main outcome of each module. The final grade will be based on the result of a mandatory assessments exercises, in the form of Multiple Choice Questions (MCQ)-like tests to complete at the end of each module.

To complete the course, the participant will have to complete a mandatory individual data-based exercise activity using R and RStudio. This assessment will be limited in time to allow time for grading.

Participation in the various chats/discussion forums will earn bonus points.

**To succeed** in the course participants must have at least a final score of 70/100, **and** comply with the following conditions:

- Participate in **all weekly webinars** and complete all mandatory activities prior to the relevant deadlines
- Complete **all weekly assessments** (tests)
- Complete the **mandatory practical exercise** with R / RStudio

**Participants must complete the end-of-course evaluation to receive a certificate.**