Reproducible Analytical Pipeline

Christophe Bontemps - SIAP[1]

15 Sep 2022

[1] *This document uses teaching materials developped by Joseph Crispell & Nathan Begbie (ONS-UK)*

Fundamental Principles of Official Statistics

- Clear mention of the **process** used to produce statistics

Fundamental Principles of Official Statistics

- Clear mention of the **process** used to produce statistics
- To retain trust in official statistics, the statistical agencies need to decide according to strictly professional considerations, including scientific principles and professional ethics, on the **methods and procedures** for the collection, **processing**, storage and presentation of statistical data.

Usual practice: Theory *vs* reality

Usual practice: In the end

What are the issues?

- Lots of files
- Cut and paste is not a reliable, reproducible approach!
- Each operator has his/her own approach
- Several versions of code may coexist
- Mistakes hard to track
- The steps aren’t recorded
- Testing is hard
- Reproducibility is not granted
- Quality is controlled only at the end

What is a Reproducible Analytical Pipeline (RAP)?

- It is a process
- It is easily repeatable
- It is easily extendable
- It is automated
- It minimises mistakes
- It is fast
- It builds trust

Image taken from: The Turing Way book

What does a RAP look like?

It is a simple process:

- linking **inputs** (data)
- to **outputs** (publication)

What does a RAP look like?

What does a RAP look like?

All the functionality in one place

What are the benefits?

Analysis within an RAP are:

- Easy to use
- Easy to find information
- Easy for others to use
- Easy to revise and adapt
- Easy to reuse
- Automated and fast
- Open and promoting trust

Image taken from: The Turing Way book

What do we need?

- A good organisation
  - of files
  - of code
  - of documentation
- Open source software
- A versioning system
- Time to learn

Why open-source instead of proprietary?

Open source tools are:

- Used by millions - huge supportive online community
- Flexible to all data sources
- Free for anyone to use - it is easier to share
- Flexible to all output types

Image taken from: The Turing Way book

What is version control?

Tracking the three **W**s: **W**ho made **W**hich change and **W**hy?

For example, take a look at this presentation

Why use version control?

- One place to store your code
- You and collaborators are free to write and develop locally
- Complete documented history of all changes made
- Easy to share
- Your future self will thank you!

The 4 Rs!

An analysis can be:

- Reproducible
- Replicable
- Robust
- Reusable

Image adapted from: The Turing Way book

What do we mean by reproducible?

> *A project is reproducible if it returns the same results when redone with the same data and the same analysis (same code).*

What are the benefits?

- Helps build trust
- Not reliant on single individual
- Can be adapted and re-used

Image adapted from: The Turing Way book

Making a RAP is difficult

Before we start, here are a few things to consider:

- IT infrastructure available
- Data privacy - where and how am I storing my data?
- Expertise - what training do I need?
- Legacy systems - what are the barriers to transitioning?

But it is worth it!

Image taken from: The Turing Way book

And we don't have to do it all at once nor alone

The building blocks of a RAP:

- Version control
- Using open-source tools
- Create reproducible code

… are useful in their own right, each will improve a specific dimension of work.

RAP in practice –> Vanuatu?

RAP has been successfully rolled out in 10s of projects across the UK government.

It is now part of the best practice documentation.

Work continues across the government to roll out RAP to more projects.

Useful resources

- The UK government RAP website.
- A free RAP course to teach you all you need to know.
- How the Data Science Campus sets its coding standards.
- A new open-source book from the Alan Turing institute setting out how to do reproducible data science.
- The github page for this presentation and other materials is here.

Citing *The Turing Way*

Many of the beautiful images used in this presentation were taken from *The Turing Way* book.

Full citation:

The Turing Way Community, Becky Arnold, Louise Bowler, Sarah Gibson, Patricia Herterich, Rosie Higman, . . . Kirstie Whitaker. (2019, March 25). The Turing Way: A Handbook for Reproducible Data Science (Version v0.0.4). Zenodo. http://doi.org/10.5281/zenodo.3233986