Dashboards project

Richard White

2018-11-01

Contents

List of Tables

List of Figures

Preface

The dashboards project is a project at FHI concerned with running automated analyses on data. An automated analysis is any analysis that:

1. Will be repeated multiple times in the future
2. Always has an input dataset with consistent file structure
3. Always has the same expected output (e.g. tables, graphs, reports)

1 Introduction

1.1 Executive summary

The dashboards project is a project at FHI concerned with running automated analyses on data.

In principle, the dashboards project is split up into two parts:

1. The umbrella infrastructure (i.e. Docker containers, continuous integration, chron jobs, etc.)
2. The R package for each automated analysis

1.2 What is an automated analysis?

An automated analysis is any analysis that:

1. Will be repeated multiple times in the future
2. Always has an input dataset with consistent file structure
3. Always has the same expected output (e.g. tables, graphs, reports)

1.3 Why not have one project for each automated analysis?

Automated analyses have a lot of code and infrastructure in common.

Automated analyses:

1. Need their code to be tested via unit testing to ensure the results are correct
2. Need their code to be tested via integration testing to ensure everything runs
3. Need to be run at certain times
4. Need to be able to send emails notifying people that the analyses have finished running
5. Need to make their results accessible to the relevant people

By combining them all in one umbrella project we can force everyone to use the same infrastructure and coding principles, so we:

1. Only need to solve a problem once
2. Only need to maintain one system
3. Can easily work on multiple projects, as we all speak the same language

1.4 Important repositories

1.4.1 Infrastructure

https://github.com/raubreywhite/dashboards_control/ (private)

This contains the Dockerfiles, cronfiles, all bash scripts, etc.

http://github.com/raubreywhite/docker/

This contains the base analysis Dockerfile

https://rocker-project.org

https://folkehelseinstituttet.github.io/fhi/

This is an R package that contains helper functions.

1.4.2 Automated analyses R packages

https://folkehelseinstituttet.github.io/dashboards_sykdomspuls/

https://folkehelseinstituttet.github.io/dashboards_normomo/

https://folkehelseinstituttet.github.io/dashboards_sykdomspuls_pdf/

https://folkehelseinstituttet.github.io/dashboards_noispiah/

https://folkehelseinstituttet.github.io/dashboards_sykdomspuls_log/

2 Umbrella Infrastructure

2.1 Physical Hardware and Subscriptions

- One Github organization (http://github.com/folkehelseinstituttet/)
- One Github team (https://github.com/orgs/folkehelseinstituttet/teams/dashboards)
- One drat repository (https://folkehelseinstituttet.github.io/drat/)
- One travis-ci.org account (http://travis-ci.org/folkehelseinstituttet)
- One travis-ci.com account (http://travis-ci.com/folkehelseinstituttet)
- One Docker hub account (http://hub.docker.com/u/raw996/)
- At least three computers:
    1. Production linux computer `smhb`

2. Testing linux computer `linux`
3. Development linux computers (1 per person)

2.1.1 Requirements - smhb

- Git
- Docker Engine - Community (https://www.docker.com/products/docker-engine)

2.1.2 Requirements - linux

- Git
- Docker Engine - Community (https://www.docker.com/products/docker-engine)
- Jenkins installed via a Docker container (http://jenkins.io)

2.1.3 Requirements - dev

- Git
- Docker Engine - Community (https://www.docker.com/products/docker-engine)

2.2 Analysis Docker Image

2.2.1 Images

Our analysis Docker images are based off the rocker images. More specifically, the rocker/verse:3.5.0 image.

This Docker image is then expanded upon by a separate Dockerfile raw996/dhadley. This Docker image is automatically rebuilt by `Jenkins` on `linux` whenever the repository is updated. The resultant Docker image is pushed to raw996/dhadley:3.5.0. This image is a general-purpose analysis image, with no sensitive information in it.

This Docker image is then expanded upon by a separate Dockerfile raw996/dashboards_r. This Docker image is automatically rebuilt by `Jenkins` on `linux` whenever the repository is updated. The resultant Docker image is locally tagged as `raw996/dashboards_r:test` and then a number of integration tests are performed on it. If the integration tests are passed, then the Docker image is retagged and pushed to raw996/dashboards_r:production. This image is private as it contains passwords and email addresses.

2.2.2 File Structure

Inside `raw996/dashboards_r` we have the following file structure:

```
/data_raw/
  |-- normomo/
  |-- noispiah/
  |-- sykdomspuls/
  |-- sykdomspuls_pdf/
  |-- sykdomspuls_log/
/data_clean/
  |-- normomo/
  |-- noispiah/
  |-- sykdomspuls/
  |-- sykdomspuls_pdf/
  |-- sykdomspuls_log/
/data_app/
  |-- normomo/
  |-- noispiah/
  |-- sykdomspuls/
  |-- sykdomspuls_pdf/
  |-- sykdomspuls_log/
/results/
  |-- normomo/
  |-- noispiah/
  |-- sykdomspuls/
  |-- sykdomspuls_pdf/
  |-- sykdomspuls_log/
/usr/local/lib/R/site-library/  <soft linked to /r>
  |-- <OTHER R PACKAGES INSTALLED HERE>/
  |-- fhi/
  |-- normomo/
  |-- noispiah/
  |-- sykdomspuls/
  |-- sykdomspuls_pdf/
  |-- sykdomspuls_log/
  |-- <OTHER R PACKAGES INSTALLED HERE>/
```

Note that we have a soft link between `/r` and `/usr/local/lib/R/site-library/`.

2.2.3 cron

We use cron to schedule the analyses. The schedule is specified in crontab.

The cronjobs are only activated when the environmental variable `ADD=cron` is defined. Cronjobs are then activated through add_cron.sh.

In principle, cronjobs should only be activated on `smhb`.

2.2.4 autofs

We use autofs to connect to the F network. The network locations, username, and password are specified in auto.mounts.

Autofs is only activated when the environmental variable `ADD_AUTOFS=yes` is defined. Autofs is then activated through add_autofs.sh.

In principle, autofs should only be activated on `smhb`.

2.3 Reverse Proxy Docker Image

We use nginx as a reverse proxy to make rstudio server available to the developers.

The relevant Dockerfile is [here](raw996/dashboards_r and is pushed to raw996/dashboards_nginx:product after integration testing is passed.

2.4 Docker Compose

Docker compose is used to integrate these Docker images into the local filesystem. We have multiple docker-compose files for different reasons:

- For production on `smhb`
- For testing on `linux`
- For development on a dev computer

3 R packages

3.1 Generic

3.1.1 Overview

Each automated analysis has its own R package:

- sykdomspuls
- normomo
- noispiah
- sykdomspulspdf
- sykdomspulslog

Each R package contains all of the code necessary for that automated analysis. Typical examples are:

- Data cleaning

- Signal analysis
- Graph generation
- Report generation

3.1.2 Requirements

The R packages should be developed using unit testing as implemented in the testthat package.

Furthermore, the R package should operate (and be able to be tested) independently from the real datasets on the system. This is because the real datasets cannot be shared publically or uploaded to github. To circumvent this issue, each package will need to develop functions that can generate fake data. GenFakeDataRaw is one example from sykdomspuls.

We also require that unit tests are created to test the formatting/structure of results. ValidateAnalysisResults is one example from sykdomspuls, where the names of the data.table are checked against reference values to ensure that the structure of the results are not accidentally changed.

3.1.3 Deployment via travis-ci and drat

Unit testing is then automatically run using travis-ci. If the R package passes all tests, then we use drat to deploy a built version of the package to Folkehelseinstituttet's R repository: https://folkehelseinstituttet.github.io/drat/.

3.1.4 Integration with the local file system

4 Integrating the R package into the physical system

4.1 Summary

An R package is not enough to run an analysis – something needs to physically call the functions inside the R package. That is, the R package needs to be integrated into the physical system.

Everything related to integrating the R package into the physical system lives in the dashboards repository.

Inside the dashboards repository we have:

```
- dev/
  |-- src/
    |-- sykdomspuls/
       |-- 0_run.sh
```

```
   |-- RunProcess.R
   |-- RunTest.R
|-- normomo/
   |-- 0_run.sh
   |-- RunProcess.R
   |-- RunTest.R
|-- sykdomspuls_log/
   |-- 0_run.sh
   |-- RunProcess.R
   |-- RunTest.R
|-- sykdomspuls_pdf/
   |-- 0_run.sh
   |-- RunProcess.R
   |-- RunTest.R
```

4.2 RunProcess.R

4.2.1 Aim

An automated analysis needs to:

1. Know the location of the data/results folders.
2. Check for new data in these folders. If no new data - then quit.
3. Load in the data.
4. Load in the analysis functions.
5. Run the analyses.
6. Save the results.

`RunProcess.R` is responsible for these tasks.

We can think of it as an extremely short and extremely high-level script that implements the analysis scripts.

Depending on the automated analysis `RunProcess.R` can be run every two minutes (constantly checking for new data), or once a week (when we know that data will only be available on a certain day/time).

4.2.2 Bounded context

1. Only one instance of `RunProcess.R` can be run at a time.
2. Data only exists on physical folders on the system.
3. The following folder structure exists on the system (here the name of the automated analysis is `ANALYSIS`):

`/data_raw/`

```
  |-- ANALYSIS/
/data_clean/
  |-- ANALYSIS/
/data_app/
  |-- ANALYSIS/
/results/
  |-- ANALYSIS/
/src/
  |-- ANALYSIS/
     |-- 0_run.sh
     |-- RunProcess.R
     |-- RunTest.R
```

Point #1 is important because if `RunProcess.R` is run every 2 minutes (constantly checking for new data) but the analyses take 3 hours to run, then we need to ensure that only one instance of `RunProcess.R` can be run at a time.

Point #2 is important because sometimes:

1. Data files need to be downloaded from external SFTP servers (normomo, sykdomspulslog).
2. Results files need to be uploaded to external SFTP servers (sykdomspuls).

If we include code to download/upload the files from SFTP servers inside `RunProcess.R` then it makes it very difficult to test `RunProcess.R` (because we will then need to simulate SFTP servers inside our testing infrastructure). If we know that `RunProcess.R` only accesses files that are available on physical folders in the system, then our testing infrastructure is a lot easier to create and maintain.

4.3 0_run.sh

4.3.1 Aim

The aim of `0_run.sh` is to ensure that:

1. Points 1 and 2 of the bounded context of `RunProcess.R` happen
2. Run `RunProcess.R`

With regards to the bounded context, we ensure that only one instance of `RunProcess.R` is run at a time through the use of `flock`.

(If neccessary) with regards to the bounded context, we use `sshpass`, `sftp`, and `ncftpput` to download/upload files from SFTP servers.

We then run `RunProcess.R` with a standard call:

```
/usr/local/bin/Rscript /src/ANALYSIS/RunProcess.R
```

4.4 RunTest.R

4.4.1 Aim

The aim of `RunTest.R` is to perform integration testing on the automated analysis. This integration testing is performed as part of the Jenkins build pipeline.

5 Contributing

5.1 Development guidelines

We try to follow the GitHub flow for development.

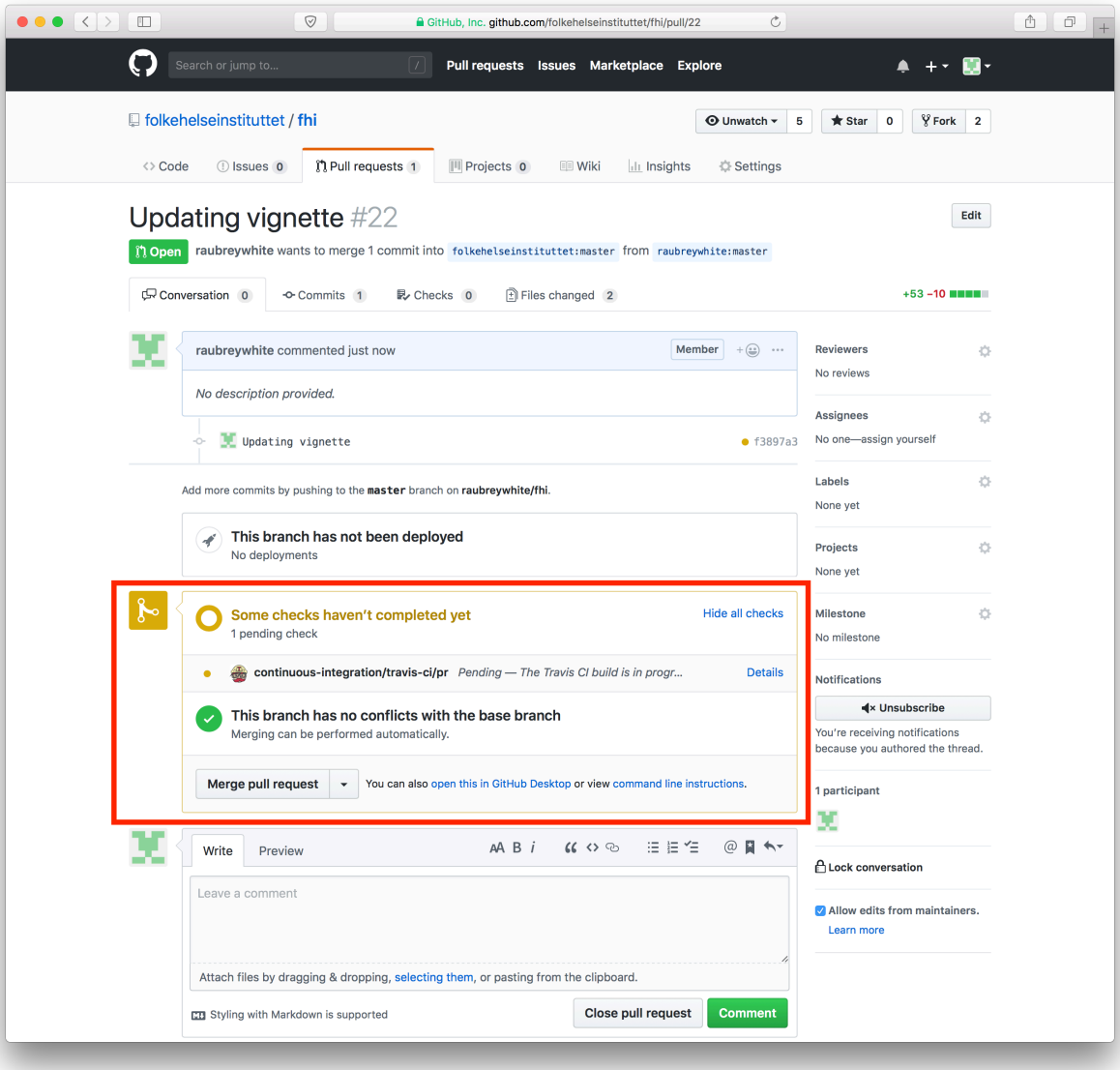
1. Fork [this repo][repo] and clone it to your computer. To learn more about this process, see this guide.
2. Add the Folkehelseinstituttet repository as your upstream:

   ```
   git remote add upstream https://github.com/folkehelseinstituttet/ORIGINAL_REPOSITOR
   ```

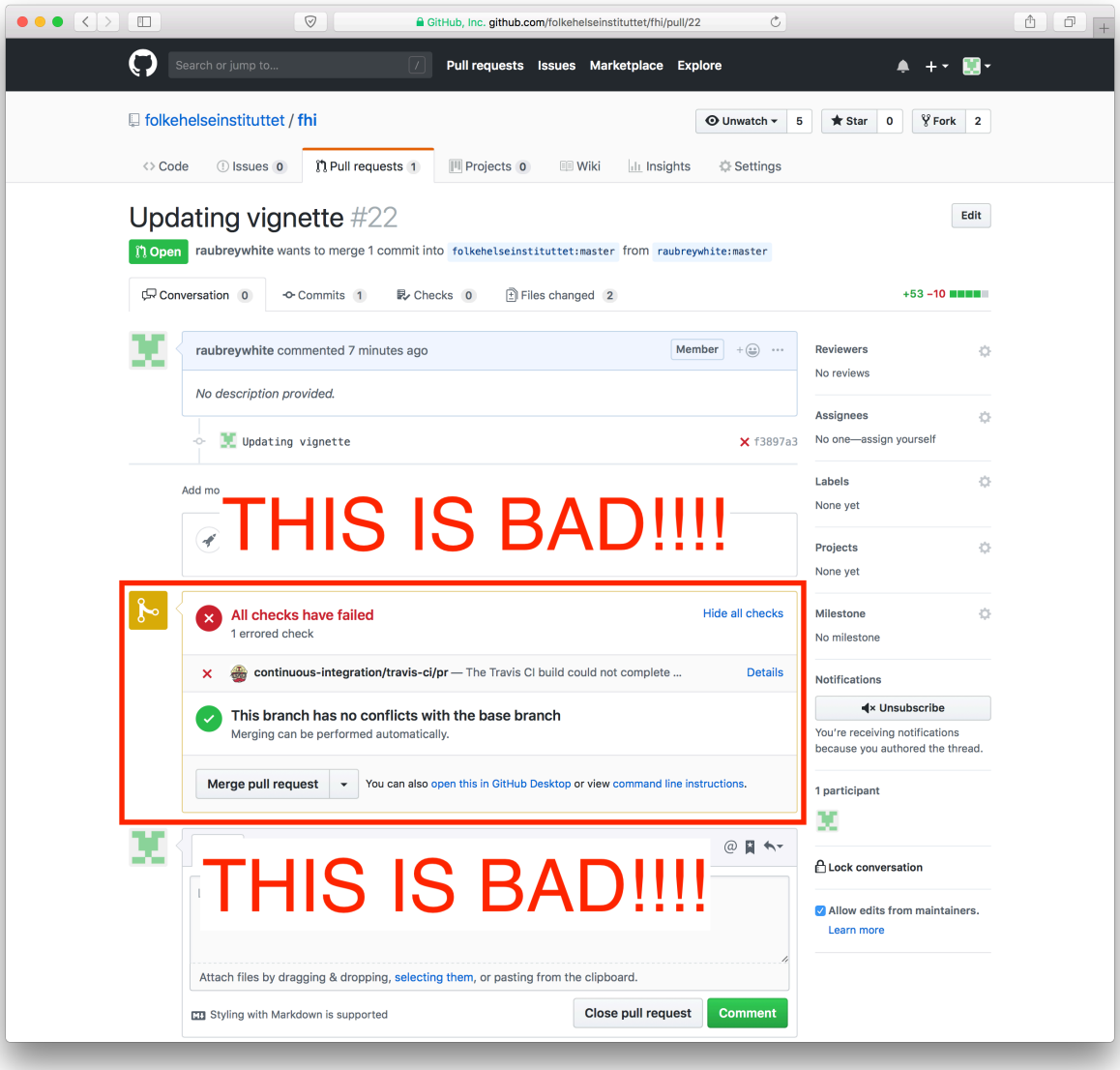
3. If you have forked and cloned the project before and it has been a while since you worked on it, merge changes from the original repo to your clone by using:

   ```
   git fetch upstream
   git merge upstream/master
   ```

4. Open the RStudio project file (`.Rproj`).
5. Make your changes:
   - Write your code.
   - Test your code (bonus points for adding unit tests).
   - Document your code (see function documentation above).
   - Do an `R CMD check` using `devtools::check()` and aim for 0 errors and warnings.
   - Commit your changes locally
   - Merge changes from the original repo (again)
   - Do an `R CMD check` using `devtools::check()` and aim for 0 errors and warnings.
6. Commit and push your changes.
7. Submit a pull request.
8. If you are reviewing the pull request, wait until the travis-ci unit tests have finished



9. Please make sure that the unit tests `PASS` before merging in!!



5.2 Code style

- Function names start with capital letters
- Variable names start with small letters
- Environments should be in ALL CAPS
- Reference Hadley's style code
- <- is preferred over = for assignment
- Indentation is with two spaces, not two or a tab. There should be no tabs in code files.

- if () {} else {} constructions should always use full curly braces even when usage seems unnecessary from a clarity perspective.
- TODO statements should be opened as GitHub issues with links to specific code files and code lines, rather than written inline.
- Follow Hadley's suggestion for aligning long functions with many arguments:

```
long_function_name <- function(a = "a long argument",
                               b = "another argument",
                               c = "another long argument") {
  # As usual code is indented by two spaces.
}
```

- Never use print() to send text to the console. Instead use message(), warning(), and error() as appropriate.
- Use environment variables, not options(), to store global arguments that are used by many or all functions.