Instructions to set-up for CARDAT (CoESRA-CAR)

Contents

Introduction

This document provides instructions for using Centre for Air pollution, energy and health Research (CAR) Data Access Technology.

CARDAT is a collection of IT infrastructure that enables easy data sharing and reuse, and reproducible data analysis.

It consists of:

- the CAR Data Inventory (a catalogue of available datasets and their associated metadata records,
- a cloud based file store (CloudStor) which enables shared access to datasets, and the ability to sync data automatically, and
- an online platform for data analysis (CoESRA) which provides a secure environment for reproducible, collaborative data analysis.

Data Platform Access and Use

Accessing the CAR Data Inventory

The CAR data inventory is used to generate a catalog of available datasets which is published to the CARES dropbox folder. Go to the folder CARES_working_temporary/Data inventory/_book and open the index.html file in FIREFOX.

This will open a web page with navigation pane for browsing data, and search bar for searching dataset (the search bar only works using Firefox).

Accessing CloudStor

CloudStor is a cloud based file sharing platform which can be accessed either through a web platform, or through a desktop application that synchronises to a local file system.

The datasets stored on CloudStor are categorized into regular data (available to everyone in CAR) and restricted data (access can be requested through the Data Manager, Ivan Hanigan).

Accessing the CAR files on CloudStor (with an AAF login)

1. For those with AAF logins there are instructions for setting up an account, and how to use both the web and desktop applications here: https://support.aarnet.edu.au/hc/en-us/articles/227469547-CloudStor-Getting-Started-Guide.
2. To gain access to the CAR datasets contact the data administrator ivan.hanigan@sydney.edu.au who will share the relevant data folders.

Accessing the CAR files on CloudStor (without an AAF login)

For those without AAF logins there are three alternative ways to be granted access to data:

1. CloudStor folders can be sent data as a password protected URL links on request to the data administrator ivan.hanigan@sydney.edu.au.
2. users can be added as guests to the AAF, or
3. datasets can be downloaded, zipped and emailed on request.

Instructions for setting up CoESRA

This is the step-by-step instructions to register for access to the Collaborative Environment for Scholarly Research and Analysis (CoESRA) portal. The CoESRA portal provides researchers with services to share and process research data on the cloud and maintain access control to ensure licence restrictions are met. This infrastructure means that the whole team (including international collaborators) can work together while residing in different states/countries.

Step 1: Sign in

Go to the URL https://coesra.tern.org.au PLEASE ONLY USE GOOGLE CHROME WEB BROWSER AS OTHER BROWSERS ARE NOT SUPPORTED AND MAY NOT WORK AS EXPECTED.

CoESRA was developed for linkage with the Australian Access Federation (AAF) authentication, once registered, user will access CoESRA using AAF authentication. Alternatively access can be granted through a google account if your organisation is not part of AAF, or a guest account can be created by contacting esupport@tern.org.au.

In order to login use the login button at the top right of the landing page (outlined in red in Figure 1 below):
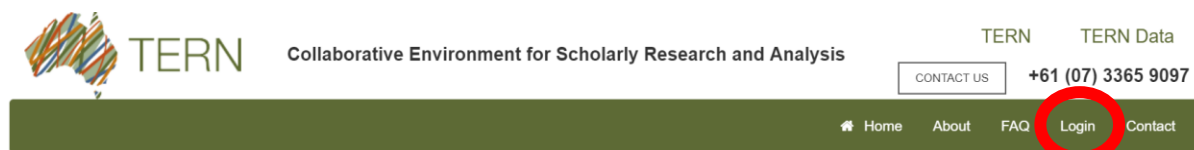

Figure 1: Logging into to CoESRA

Select between using your AAF or Google login.

Figure 2: Select how to sign into CoESRA
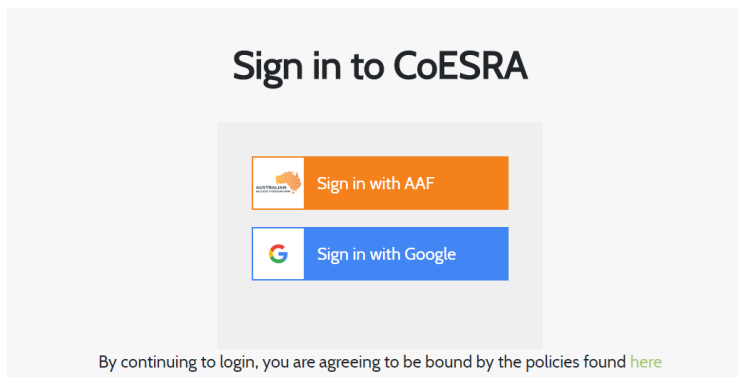
If using AAF search for and select your institution. You can check the 'Remember my organisation' checkbox to save having to search for your institution in the future. See Figure 3 as an example.

Figure 3: Searching for your institution when logging in with AAF

Complete your login using your institutional login and password.

Step 2: Request access to the shared data group

Now that you are logged in you want to ask for permission to access the shared data folder. To access the CAR environment general data (i.e. non-restricted data), go to the groups tab (or use the URL https://coesra.tern.org.au/#/group-manager).

Select the car group, and click on the Request button.

The shared data will also be available on CloudStor – for instructions on how to access the data see the section on Accessing CloudStor.

If there is restricted data you want access to contact the project manager, ivan.hanigan@sydney.edu.au to request access.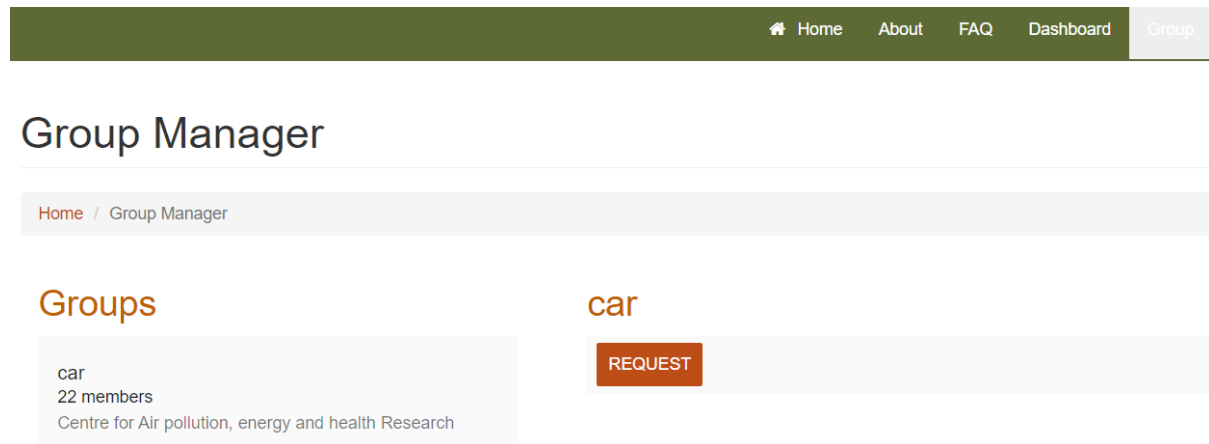 Approval times may vary if third party permissions are required. After approval, you will be granted access to the shared datasets.

Step 3: Start your remote desktop

To start a remote desktop click on the 'Launch a desktop' button (see Figure 4). Your virtual desktop will be allocated to you for a set amount of time (currently 5 days), and the expiry will be presented in a pop-up box. Click Yes.

After a short time, your new desktop should be available to select under the Running desktops heading. Click on the 'Go to desktop button'. See Figure 4.
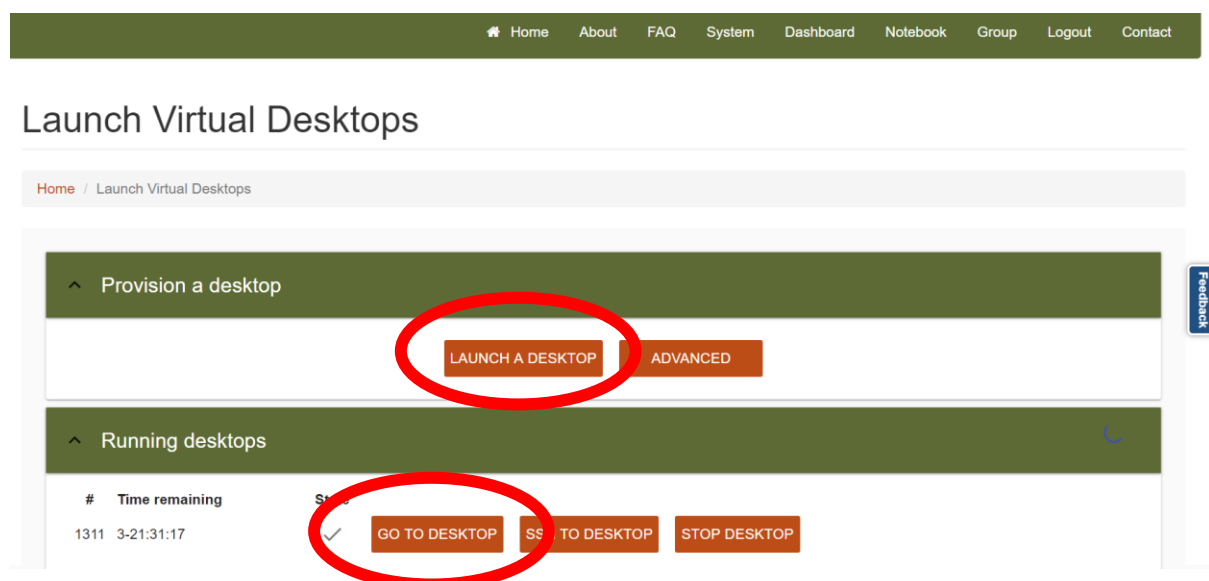

Figure 4: Launching a new desktop

There are some options for customising desktops availible by using the 'Advanced' button next to the 'Launch a desktop' button (see Figure 4). For more information refer to the 'CoESRA Features' section of the webpage https://coesra.tern.org.au/#/tern-landingpage.

You now have a folder that is just for you called "Your_Username's Home". This appears as an icon under the "Computer" icon at the top left. Double click on "Your_Username's Home" to open it. See Figure 5.
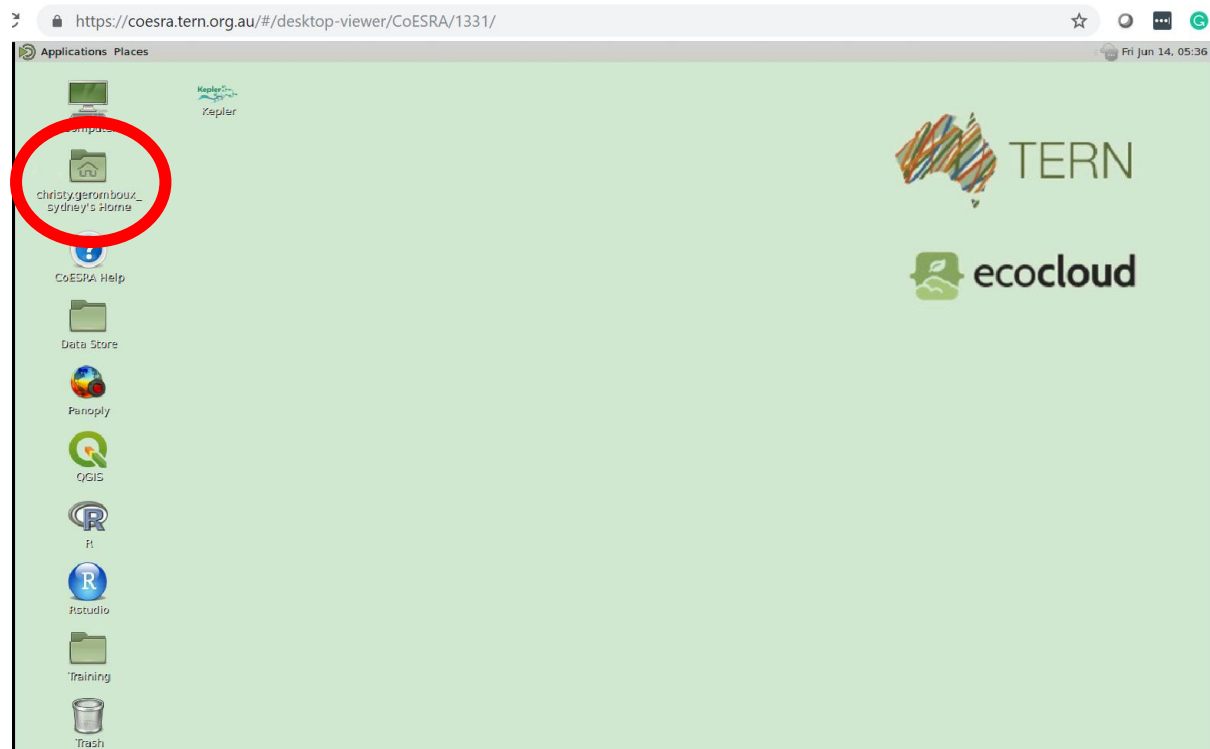

Figure 5: The personal user workspace

Then double click on the "public_share_data" link.

Now double click on the "ResearchData_CAR" folder and then the folder "Environment_General". (This is only accessible by the members of the group.)

Step 4: Install the recommended R tools and create a shortcut to the shared data

Open the Rstudio software (there is a link on the desktop, or this is under the menu 'Applications > CoESRA') and run the following codes:

```
install.packages("rgeos")
install.packages("rgdal")
install.packages("raster")
install.packages("foreign")
install.packages("sqldf")
install.packages("dplyr")
install.packages("data.table")
install.packages("reshape")
```

Step 5: Uploading data

For datasets that are not yet saved to CloudStor, the recommended approach for uploading data is to use a web service accessed from within the CoESRA virtual desktop. The CloudStor via the AARNET consortium of Australian Research Organisations is preferred as it is secure and fast, however other services such as Microsoft Onedrive, Dropbox or Google drive are also feasible alternatives.

Case Study - Instructions to run a shared workflow, and extend this with new data or methods

This is the step-by-step instructions to run the example workflow to compute a DPSEEA "package" of EH Indicators (for example this might be an assessment of premature mortality attributable to anthropogenic $PM_{2.5}$).

Step 1: Data discovery

Using the data inventory report described above the user can discover a pre-written EH indicator set and workflow as pre-written codes and data inputs.

Step 2: Data acquisition

Following the link in the 'additional metadata' field will take the user to a webpage that displays the contents of the shared analysis workflow folder, as shown below. Copying the path will enable the user to easily find the location of that shared folder.
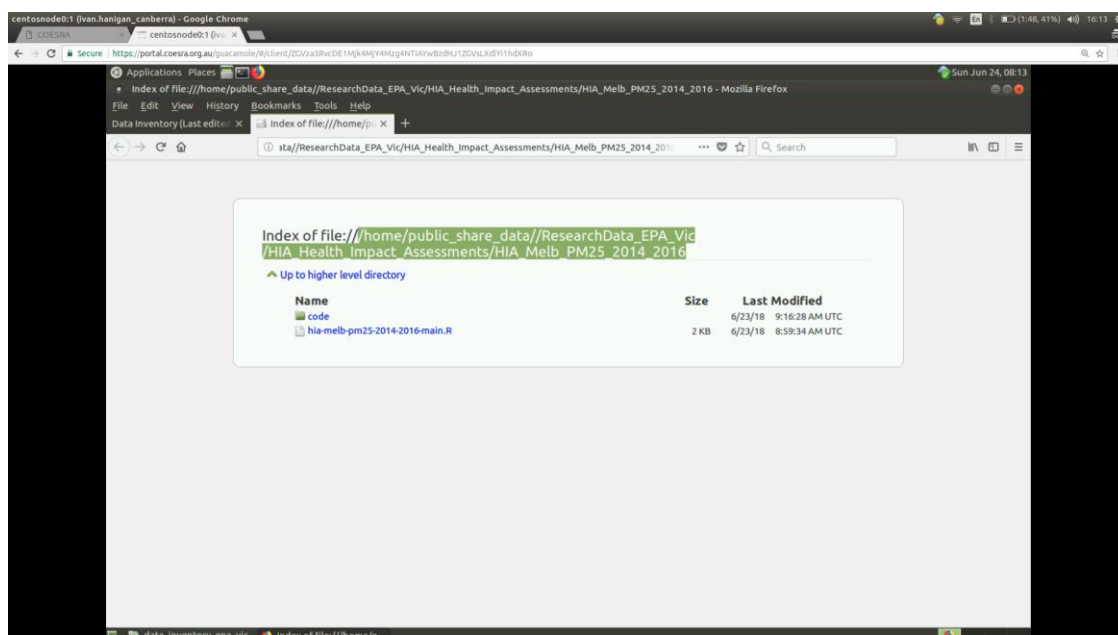

Figure 6: The data acquisition step starts with copying the path

This path can be copied into the folder browser and the user will have access to the files. This will be copied and pasted into their own home folder (the shared folder will not allow general users to write new files into this folder).
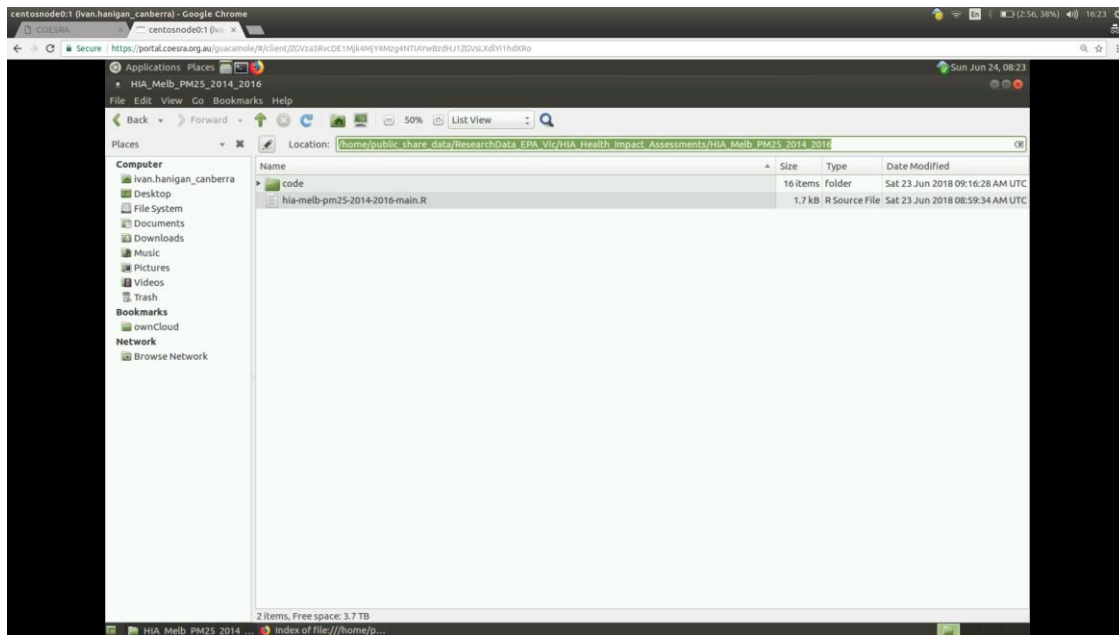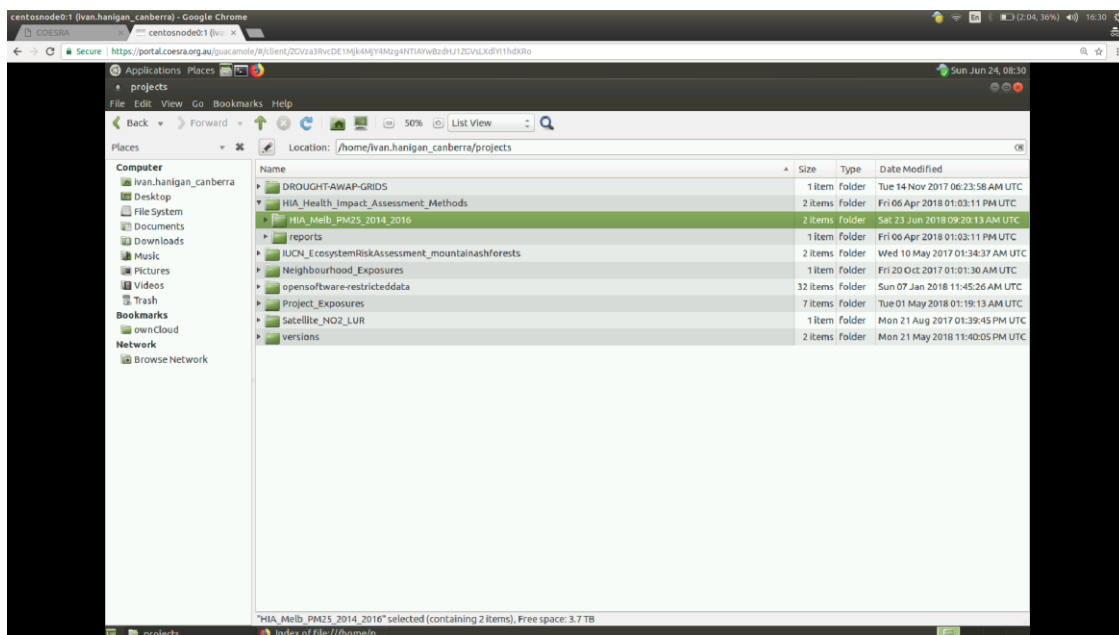
Figure 7: The shared folder

Figure 8: Paste the workflow folder into the user's home folder

Step 3: Data analysis

At the top level directory for each dataset we have an .Rproj file which we use to define the working directory, and that lists all the R scripts that can be run and what they do.

Older code used Main R Script at the top level directory for the analysis. For these you will need to load the Main R Script in RStudio and make some modifications to the key paths at the top.
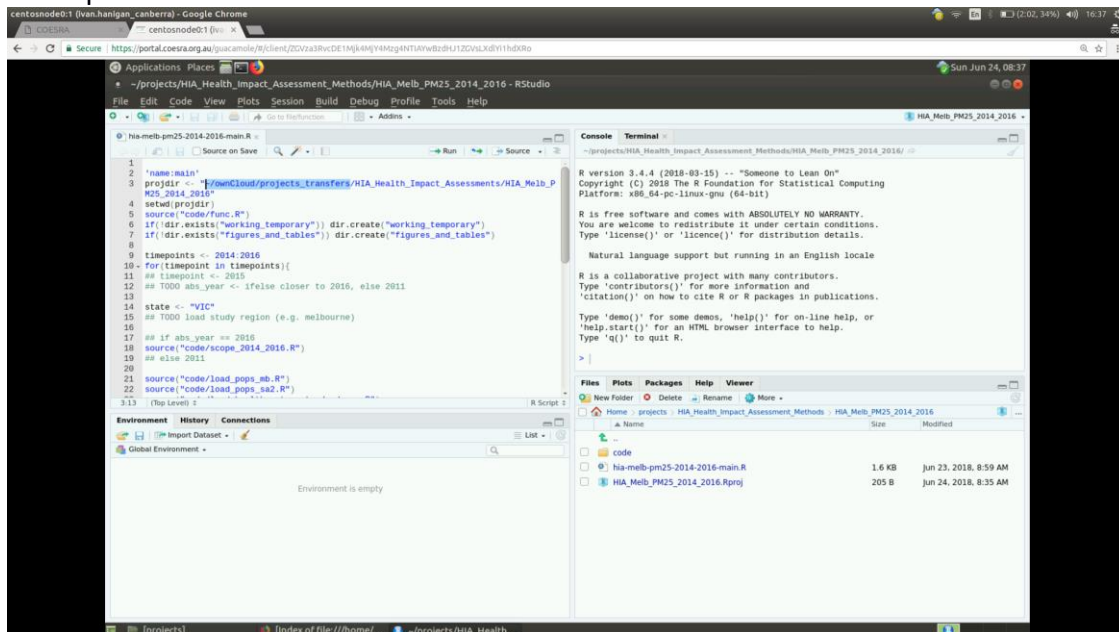

Figure 9: Make some minor modifications to the Main R Script, so that it is specific to this user

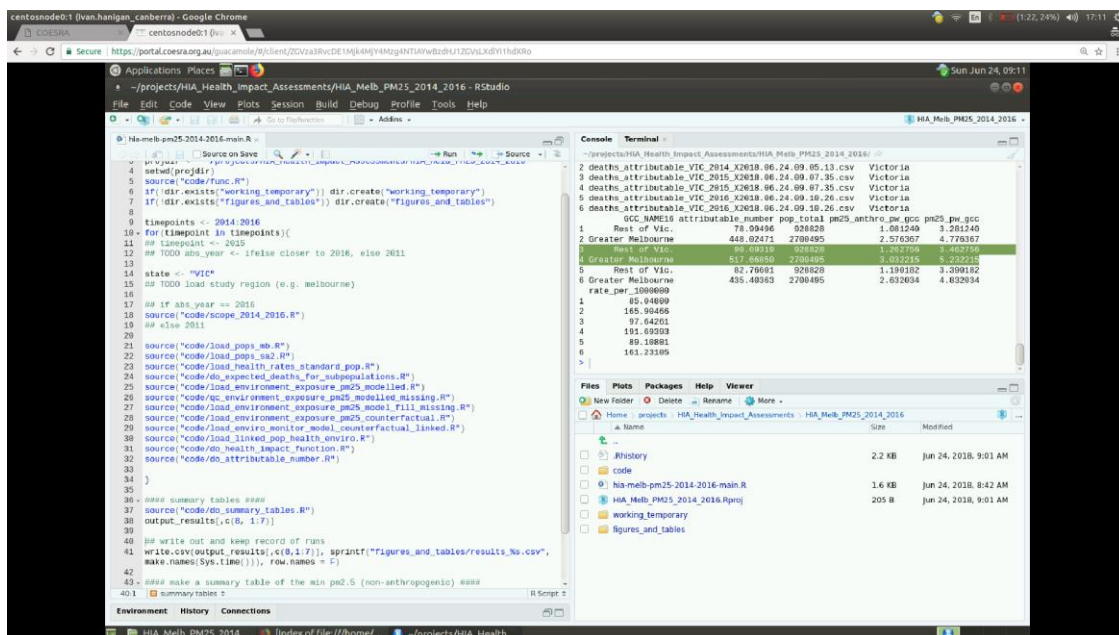At this point the shared workflow can be re-run but this user to recalculate the result as previously found.

Figure 10: The workflow can be run to compute the results and reproduce the finding

Further reading

More comprehensive information on this project can be found in:

- Guru, S., Hanigan, I.C., Nguyen, H.A., Burns, E., Stein, J., Blanchard, W., Lindenmayer, D., Clancy, T., 2016. Development of a cloud-based platform for reproducible science: A case study of an IUCN Red List of Ecosystems Assessment. Ecol. Inform. 36, 221–230. https://doi.org/10.1016/j.ecoinf.2016.08.003 - peer review journal article on CoESRA's design and implementation.
- The CAR and LHI Procedures Manual – an internal document for data curation and IT set up.