Documentation for Publication Venue Gap Analysis

This document summarises the various code involved in the data preparation and analysis for the Open Access Publication Venue Gap analysis. The code was written by Tom Kenny.

# Setting up the project folder

* The first step was/ is to create an RStudio project file (see [blog on this here](https://martinctc.github.io/blog/rstudio-projects-and-working-directories-a-beginner's-guide/)), along with subfolders for
  + ‘Data’ = this is where all the data will be stored, it should also contain a subfolder called ‘Raw data’
  + ‘Outputs’ = this is where all the tables and charts will go
  + ‘Scripts’ = this is where all the code goes

# Data scripts

The first three scripts download and clean the data ready for analysis. These scripts were last run on the week commencing 29th March 2021, with the final merged dataset then uploaded to the DataHub

## SHERPA RoMEO API

* This code downloads data on journal OA policies from the SHERPA API in .json and then filters it to return only the variables we are interested in, and converts it into a data frame with one row for each journal policy (i.e. if a journal has four policies it will have four rows). It also does a bit of basic cleaning of the data. It then saves the file out as .Rda and .xlsx to be used in subsequent parts of the code.

## Dimensions import and cleaning

* This code imports Dimensions data and cleans it. The data was downloaded for us by Katie Shamash at Dimensions as this is the easiest option, but it could also be downloaded using the Dimensions API (see code for doing this in superceded folder).
* Once the data is imported, it gets rid of duplicates, cleans variables, derives new variables, then saves it out as .xlsx and .Rda for use by future code.

## Preparing and merging PVGA data

* This data imports previously produced data from Sherpa and Dimensions, and imports new data from ESAC. It then merges them all together and creates new variables from the merged data, in particular focusing on compliance with various policy scenarios.

## Uploading final merged\_pvga to Datahub

* This code was used to upload the data created above to the datahub to allow reproducibility if anyone wants to approach this data again. Sherpa, Dimensions and ESAC all regularly update data (even of historic records), so findings do change over time. The data was updated and uploaded w/c 29th March 2021

# Analysis scripts

The other scripts cover the main analysis and various different smaller bits of analysis. These scripts were last run on the week commencing 29th March 2021, with the final merged dataset then uploaded to the DataHub

## SHERPA RoMEO API