Project Update – 29/07/2021

# To do’s

* Review and clean the grant data
* Finalise database schema, correcting issues and typos.
* Start organising data by the schema
* Create a basic RShiny app
* Carry out two/three basic analysis based on the already collected data.

## Deferred To Dos:

* Prepare person data for gender prediction.
* Continue reviewing and cleaning the panel data.
* Investigate and create a table of linked grants (either similar cluster projects or continued projects)
* Search and collect data from Grants on the Web on grants and people.

# What you did and created?

* Completed panel data collection:
  + Did a preliminary clean and organisation of the data into fewer datasets of related data, although haven’t included them in the core dataset as they need including in the database scheme and need optimisation.
* Cleaned and process the core datasets into a collection of datasets organised by the final dataset scheme (Note: the schema diagram is incomplete still with some corrections needing to be made to it):
  + Cleaned the first name and ran them through the genderizer.io API. Didn’t do any post-processing on the data, to allow greater flexibility on reuse of the data.
* Started to carry out preliminary analysis on data, in particular the central grant dataset.
* Learnt to use and started playing around with R Shiny.
  + Sketched out by hand three different UI layout that could be used for the R Shiny app.
  + Started planning out app functionality.

# What decision have you made and how they were made?

* Chose to first complete analysis in a R Notebook to get an idea of different graphs and views to include in the R Shiny app
* Decided to save the core datasets as both CSVs (with individual files for each table) and as a RDS (R Data Object). The RDS allows for the data types and formats to be preserved for further analysis and use. The CSVs are a more open format that will allow use in other programs and languages other than R.
* Decided not to break the Themes data down in multiple tables with less repeated data but instead group themes by the Theme and Research Area. While more data is repeated it keeps the complexity of the datasets down. They are not large datasets, so the memory impact is acceptable.
* Decided to limit the amount of analysis conducted on the gender data, instead keeping it in a raw form:
  + Didn’t decide if some users were non-binary or investigate and correct incorrect labels in the data, as planned, instead leaving it in the form that it was returned from the API.
  + This will allow future users of the data to interpret the data more flexibly.
  + It also saves time as this was going to be a time-consuming task to complete.

# Relevant References

|  |  |  |
| --- | --- | --- |
| **Name** | **Link** | **Notes** |
| Advanced R | http://adv-r.had.co.nz/Exceptions-Debugging.html#condition-handling | Used to develop error handling functionality |
| Stack Overflow: Advanced TryCatch statements | <https://stackoverflow.com/questions/12193779/how-to-write-trycatch-in-r> | Used to develop complex error handling functionality |
| GtR-1-API-v3.3 | <https://gtr.ukri.org/resources/GtR-1-API-v3.3.pdf> | Used to develop API access expressions |
| Progress Package Readme | https://github.com/r-lib/progress#readme | Used vinettes/tutorial to develop progress bar functionality |
| Rvest | https://rvest.tidyverse.org/ | Used to develop scaping from HTML documents functionality |
| JSONLITE | <https://cran.r-project.org/web/packages/jsonlite/vignettes/json-aaquickstart.html> | Used to request and convert JSON files into dataframes |
| Subset list based on a filter | https://stackoverflow.com/questions/6941506/subset-elements-in-a-list-based-on-a-logical-condition | Used to filter the elements of a list based on a condition and explain the use of the Filter function |
| rvest Error in open.connection(x, “rb”) | https://stackoverflow.com/questions/33295686/rvest-error-in-open-connectionx-rb-timeout-was-reached | Used to solve error with html document retreval and develop more stable method. |
| [How to convert variable (object) name into String](https://stackoverflow.com/questions/14577412/how-to-convert-variable-object-name-into-string) | https://stackoverflow.com/questions/14577412/how-to-convert-variable-object-name-into-string | Used in logging functionality to return the variable name parsed into a function |
| Regex101 | https://regex101.com/ | Used to develop regex expressions |
| DrawSQL | <https://drawsql.app/> | Used to create database schema |
| Nas Introduced by Coercion | <https://statisticsglobe.com/warning-message-nas-introduced-by-coercion-in-r> | Used to solve an issue when coercing data to the correct data type where lots of warnings would be thrown. |
| Remove periods/dots in entire data frame | https://stackoverflow.com/questions/55026965/remove-periods-dots-in-entire-data-frame | Used to develop method to remove thousand seperators from values in dataframes |
| Change column order | <https://dplyr.tidyverse.org/reference/relocate.html> | Used to have a more elegant method of reordering columns than using a select |
| Remove all special characters from a string in R | https://intellipaat.com/community/15034/remove-all-special-characters-from-a-string-in-r | Used to clean first names from person dataframe |
| genderizeR | https://github.com/kalimu/genderizeR | Package used to access the genderize API easily and process names - detailed citation available on the page |
| R for Data Science - Data Import - Parsering a vector | https://r4ds.had.co.nz/data-import.html#parsing-a-vector | Used to solve an encoding issues with names |
| How to plot multiple data series in GGPLOT | https://www.sixhat.net/how-to-plot-multpile-data-series-with-ggplot.html | Used to plot two data series to the same graph and have a legend show |
| Relational Data | https://r4ds.had.co.nz/relational-data.html | Used to plan out dataset schema and idenify appropriate primary and foreign keys |
| GGPlot Legend Title, Position and Labels | <https://www.datanovia.com/en/blog/ggplot-legend-title-position-and-labels/> | Used to customise plot legends |

# To do items for coming weeks

* To be filled out in meeting with supervisior