The Journal of Open Source Software

# DataRepExp: a R shiny Application that makes Data FAIR for Data Repositories

**Rory Chen** [1], **Vibeke S Catts**[1], **Ashleigh Vella**[1], **Juan Carlo San Jose**[2], **Sarah Bauermeister**[3,4], **Josh Bauermeister**[3,4], **Emma Squires**[3,5], **Simon Thompson**[3,5], **John Gallacher**[3,4], **and Perminder S. Sachdev**[1]

**1** Centre for Healthy Brain Ageing, Discipline of Psychiatry & Mental Health, School of Clinical Medicine, University of New South Wales, Sydney, Australia **2** Research Technology Services, University of New South Wales, Sydney, Australia **3** Dementias Platform UK, Oxford, United Kingdom **4** Department of Psychiatry, University of Oxford, Oxford, United Kingdom **5** Population Data Science, Swansea University, Swansea, United Kingdom

**DOI:** 10.xxxxxx/draft

**Software**

- Review
- Repository
- Archive

**Editor:**

**Submitted:** 01 March 2024
**Published:** unpublished



## Summary

The Data Repository Explorer, DataRepExp, is an open-source R Shiny application developed to improve the findability, accessibility, interoperability, and reusability (FAIR) (Wilkinson et al., 2016) of research data held in a data repository. The application displays standardised metadata across multiple studies including data availability by categories (such as demographics, medical history, imaging data and genomic data) to allow high-level comparison. It enables users to explore and run preliminary analysis from participants that match certain criteria. In addition, it provides features to export reports and aggregated results for data access application purposes. The application was initially developed for a discipline-specific data-sharing platform, the Dementias Platform Australia (DPAU) (CHeBA (Centre for Healthy Brain Ageing), UNSW, 2024a). Envisioning this work could be utilized by other data repositories in diverse disciplines, this demo application was created using simulated health-related data for demonstration purposes.

- Source Code: https://github.com/RoryChenXY/DataRepExp_public
- Web Application: https://rorychenxy.shinyapps.io/DataRepExp/
- Contact: xinyue.chen1@unsw.edu.au

## Statement of Need

Data repositories have become increasingly important in recent years as more emphasis has been placed on open science practices and data sharing. By making data publicly available through repositories, researchers can ensure data persistence and support data preservation, as well as facilitate the reuse of their data, thereby increasing the potential for new scientific discoveries.

However, challenges exist for data findability, accessibility, interoperability, and reusability (FAIR) (Wilkinson et al., 2016). Even though most data repositories have adopted various metadata schemas to describe the dataset (Contaxis et al., 2022), it is increasingly a challenge for researchers to find relevant data that meet research interests or needs (Gregory, 2018). For multi-study research, applying to access different datasets usually comes with diverse and complicated data-sharing requirements and workflows, extensive administrative workloads and waiting periods. Upon approval, substantial efforts of data harmonization are usually required due to inconsistent data structures and labeling conventions, and harmonised dataset are hardly reused. We found that many data repositories do not provide comprehensive metadata, nor

centralised tables for comparison. With repositories that provide data visualisation, Power BI and Tableau are commonly used but costs are incurred. R-shiny could provide more flexibility and functions at a fraction of the cost.

Designed to enable easier access to research data held in data repositories, DPAU (CHeBA (Centre for Healthy Brain Ageing), UNSW, 2024a) seeks to address these challenges with R-Shiny (Chang et al., 2023). The application designed for DPAU includes rich metadata and a set of commonly used variables (Bauermeister, Phatak, et al., 2023), identified as being of broad interest to dementia research, harmonised using the C-Surv data model (Bauermeister, Bauermeister, et al., 2023), which has been developed by Dementias Platform UK (DPUK) (Bauermeister et al., 2020), and adopted by Alzheimer's Disease Data Initiative (ADDI) (*Alzheimer's Disease Data Initiative*, 2024) and DPAU (CHeBA (Centre for Healthy Brain Ageing), UNSW, 2024a). Researchers can identify data points from participants that match certain criteria, using filters at study and/or participant levels, then explore and conduct preliminary analysis on the filtered dataset. The application also allows users to export reports and aggregated results. The exported reports can then be used when submitting a single centralised data access application form for accessing data from multiple studies through the DPAU Data Portal (CHeBA (Centre for Healthy Brain Ageing), UNSW, 2024b).

DataRepExp was created with simulated data and a list of generalized health-related variables. This work can be modified and utilized by other data repositories by adopting the discipline-specific metadata schema and common variables. Considering some repositories may hold highly sensitive data, or individual-level data may not be not available, a metadata-only version DataRepExp has also been developed, and relevant code is included in the GitHub Repository.

With rich metadata for findability, the interactive visualization dashboard for accessibility, standardization and harmonization for data interoperability and reusability, this tool can improve the FAIR(Wilkinson et al., 2016) of research data held in a data repository. R programming skill is required for reproducibility, detailed documentation and syntax is open-source and publicly available.

# Methods

DataRepExp was written using R (R Core Team, 2023) and JavaScript using the following packages:

- Shiny: shiny (Chang et al., 2023), shinydashboard (Chang & Borges Ribeiro, 2021), shinyWidgets (Perrier et al., 2024), shinyjs (Attali, 2021).
- Data manipulation: dplyr (Wickham, François, et al., 2023), tidyr (Wickham et al., 2024), tidyverse (Wickham, 2023b), forcats (Wickham, 2023a), useful (Lander, 2023), magrittr (Bache & Wickham, 2022), purrr (Wickham & Henry, 2023).
- Data Report and Visualisation: ggplot2 (Wickham, Chang, et al., 2023), plotly (Sievert et al., 2024), scales (Wickham, Pedersen, et al., 2023), DT (Xie et al., 2024), htmltools (Cheng et al., 2023), fontawesome (Iannone, 2023).

## Deployment

The Data Repository Explorer, DataRepExp, is hosted through easy-to-use shinyapps.io, while the DPAU version is hosted on AWS environment using Shiny Server for high availability, scalability, security, and compliance.

# Overview

The application layout features a side menu, through which the users can navigate through tabs, and the main view which displays the content of the selected tab.

The Journal of Open Source Software

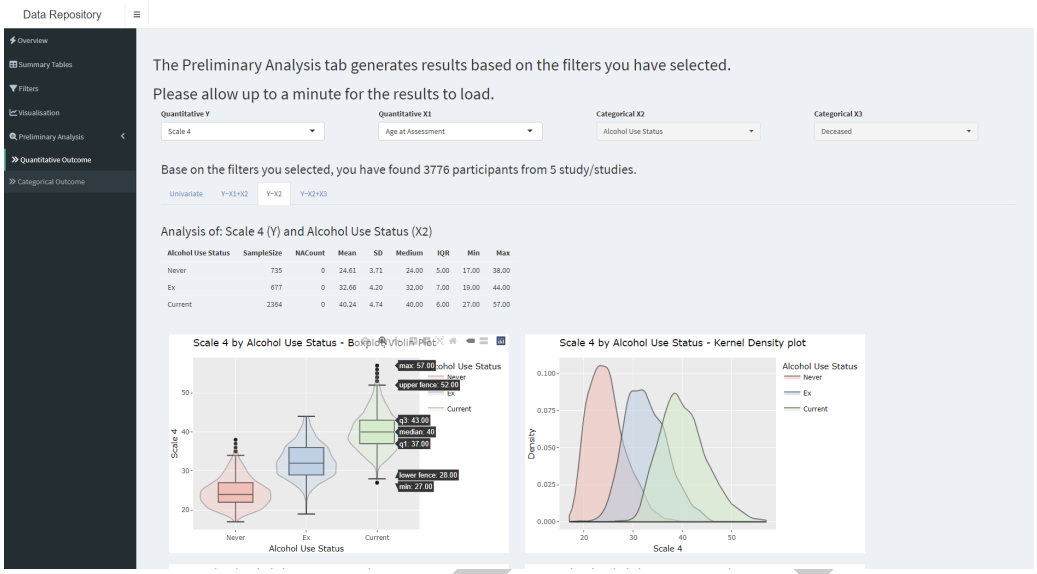


**Figure 1:** Screenshot of DataRepExp

- First tab – Overview: includes statement and navigation instructions.
- Second tab – Summary Tables: three metadata tables for high-level comparison
- Third tab – Filters and Filter Reports: users can adjust and apply filters to identify participants and studies that match selected criteria. They can download the Filter Report with the list of studies that matches the filters selected.
- Fourth tab – Visualisation: plots are organised by different domains, generated using the filtered dataset.
- Fifth tab – Preliminary Analysis: Preliminary Analysis: preliminary analysis can be done with user-selected variables.

Application features include:

- Simulation: For demonstration purposes, we generated simulated data. Scripts and reference documents used to generate the data can be found in the GitHub repository.
- Modularisation: DataRepExp was built in Shiny modules. Modularity makes the app easy to test, maintain, and deploy. The features can be easily further expanded with loose coupling module design.
- Interactive: DataRepExp provides an interactive interface that allows users to engage with the data and output. Elevated user experience with integrative charts and figures, which include functions such as sort, filter, zoom, select, adjust axis, hover for information, reset, etc.



## Acknowledgements

This application was inspired by the visualisation tool developed by DPUK (Bauermeister et al., 2020) using PowerBI, then developed in R (R Core Team, 2023) for the DPAU (CHeBA (Centre for Healthy Brain Ageing), UNSW, 2024a). We acknowledge the generous sharing of best practices and knowledge from DPUK.

## Funding

This work is supported by grants from the National Institute on Aging/ National Institute of Health (NIA/NIH) [1RF1AG057531-01] and the Medical Research Council [MRC/T0333771].

## Availability and Community Guidelines

The application and source code are available at the GitHub repository. Users and contributors are welcome to contribute, request features, and report bugs through the GitHub repository.

## References


*Alzheimer's Disease Data Initiative*. (2024). https://www.alzheimersdata.org/

Attali, D. (2021). *Shinyjs: Easily improve the user experience of your shiny apps in seconds*. https://deanattali.com/shinyjs/

Bache, S. M., & Wickham, H. (2022). *Magrittr: A forward-pipe operator for r*. https://magrittr.tidyverse.org

Bauermeister, S., Bauermeister, J. R., Bridgman, R., Felici, C., Newbury, M., North, L., Orton, C., Squires, E., Thompson, S., Young, S., & others. (2023). Ready data: The c-surv data model. *European Journal of Epidemiology*, *38*(2), 179–187.

Bauermeister, S., Orton, C., Thompson, S., Barker, R. A., Bauermeister, J. R., Ben-Shlomo, Y., Brayne, C., Burn, D., Campbell, A., Calvin, C., & others. (2020). The dementias platform UK (DPUK) data portal. *European Journal of Epidemiology*, *35*, 601–611.

Bauermeister, S., Phatak, M., Sparks, K., Sargent, L., Griswold, M., McHugh, C., Nalls, M., Young, S., Bauermeister, J., Elliott, P., & others. (2023). Evaluating the harmonisation potential of diverse cohort datasets. *European Journal of Epidemiology*, *38*(6), 605–615.

Chang, W., & Borges Ribeiro, B. (2021). *Shinydashboard: Create dashboards with shiny*. http://rstudio.github.io/shinydashboard/

Chang, W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A., & Borges, B. (2023). *Shiny: Web application framework for r*. https://shiny.posit.co/

CHeBA (Centre for Healthy Brain Ageing), UNSW. (2024a). *Dementias Platform Australia*. https://www.dementiasplatform.com.au/

CHeBA (Centre for Healthy Brain Ageing), UNSW. (2024b). *Dementias Platform Australia Data Portal*. https://portal.dementiasplatform.com.au/

Cheng, J., Sievert, C., Schloerke, B., Chang, W., Xie, Y., & Allen, J. (2023). *Htmltools: Tools for HTML*. https://github.com/rstudio/htmltools

Contaxis, N., Clark, J., Dellureficio, A., Gonzales, S., Mannheimer, S., Oxley, P. R., Ratajeski, M. A., Surkis, A., Yarnell, A. M., Yee, M., & others. (2022). Ten simple rules for improving research data discovery. *PLoS Computational Biology*, *18*(2), e1009768.

Gregory, S. J. A. M., Kathleen AND Khalsa. (2018). Eleven quick tips for finding research data. *PLOS Computational Biology*, *14*(4), 1–7. https://doi.org/10.1371/journal.pcbi.1006038

Iannone, R. (2023). *Fontawesome: Easily work with font awesome icons*. https://github.com/rstudio/fontawesome

Lander, J. P. (2023). *Useful: A collection of handy, useful functions*. https://github.com/jaredlander/useful

Perrier, V., Meyer, F., & Granjon, D. (2024). *shinyWidgets: Custom inputs widgets for shiny*. https://github.com/dreamRs/shinyWidgets

R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Sievert, C., Parmer, C., Hocking, T., Chamberlain, S., Ram, K., Corvellec, M., & Despouy, P. (2024). *Plotly: Create interactive web graphics via plotly.js*. https://plotly-r.com

Wickham, H. (2023a). *Forcats: Tools for working with categorical variables (factors)*. https://forcats.tidyverse.org/

Wickham, H. (2023b). *Tidyverse: Easily install and load the tidyverse*. https://tidyverse.tidyverse.org

Wickham, H., Chang, W., Henry, L., Pedersen, T. L., Takahashi, K., Wilke, C., Woo, K., Yutani, H., & Dunnington, D. (2023). *ggplot2: Create elegant data visualisations using the grammar of graphics*. https://ggplot2.tidyverse.org

Wickham, H., François, R., Henry, L., Müller, K., & Vaughan, D. (2023). *Dplyr: A grammar of data manipulation*. https://dplyr.tidyverse.org

Wickham, H., & Henry, L. (2023). *Purrr: Functional programming tools*. https://purrr.tidyverse.org/

Wickham, H., Pedersen, T. L., & Seidel, D. (2023). *Scales: Scale functions for visualization*. https://scales.r-lib.org

Wickham, H., Vaughan, D., & Girlich, M. (2024). *Tidyr: Tidy messy data*. https://tidyr.tidyverse.org

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., Silva Santos, L. B. da, Bourne, P. E., & others. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 1–9.

Xie, Y., Cheng, J., & Tan, X. (2024). *DT: A wrapper of the JavaScript library DataTables*. https://github.com/rstudio/DT