

International Journal of Population Data Science

Journal Website: www.ijpds.org



Establishing a Repository of Synthetic Datasets for Researchers: A Scottish Perspective

Sophie McCall¹

¹Research Data Scotland, Edinburgh, United Kingdom

Objectives

This presentation will describe how we co-designed the creation and provision of a single low fidelity synthetic data repository for multiple data controllers, where researchers can access synthetic data for data discovery and code development.

researcher access to data. By adopting an iterative approach and responding to user and public feedback, our repository has proved to be an exemplar of how to make synthetic data available.

Method

From user and public engagement, we identified researcher demand for access to low-fidelity synthetic data prior to real data. Our metadata catalogue collates information on Scottish datasets available for research and is the digital platform for our synthetic data repository, hosting assets generated by partner organisations and ourselves. Embedded into our process are quality checks to assess labelling, structure, disclosure and documentation of synthetic data, with an 'End User Licence Agreement' to guard against synthetic data being used inappropriately. These measures give assurance that privacy is preserved whilst making synthetic data as freely available as possible.

Results

We have completed the pilot phase of this project, establishing a repository of synthetic data in our metadata catalogue, which researchers can apply to for access. We will share the rationale for decisions made during the project, together with challenges faced. Key aspects of consideration are the user journey and promotion of our service within the wider data community. Using data analytics relating to number of synthetic datasets requested and downloaded, together with case studies from researchers, we will establish the success of our project. Finally, we will describe our plans to extend our work, to include hosting more datasets and working with partner organisations to support with their synthetic data generation, to ensure they meets our standards of quality and disclosure.

Conclusion

Developing a synthetic data repository has been a significant milestone for our organisation and our ambition to improve



Temporary page!

$\LaTeX$ was unable to guess the total number of pages correctly. As there was some unprocessed data that should have been added to the final page this extra page has been added to receive it.

If you rerun the document (without altering it) this surplus page will go away, because $\LaTeX$ now knows how many pages to expect for this document.