

International Journal of Population Data Science

Journal Website: www.ijpds.org



Developing Generalisable Linkage Methodologies for Administrative Datasets

Leah Quinn¹

¹Office for National Statistics, Manchester, United Kingdom

Objectives

Bespoke data linkage methodologies are time consuming to develop. The Generalisable Linkage of Administrative Demographic Index Service (GLADIS) project aims to minimise the need for bespoke methods by designing generalisable linkage methodologies to link a variety of administrative datasets to a population spine.

Conclusion

GLADIS will make linked data more accessible, increase the consistency and comparability of linked datasets, and improve compatibility for onward linkages. This will support the Integrated Data Service; a flagship platform to enable quick, easy access to data to facilitate analysis for the public good and inform decision making.

Methods

A pipeline module was created for each key stage in linkage projects. We will present novel methods such as automatically identifying probabilistic parameters without manual user involvement, and using record 'explosion' to efficiently account for alternative variables in clusters by creating additional derived rows for linking. Together, the GLADIS methodologies not only allow good quality linkage but also minimise the required user input and the necessary expertise of the linkage analyst. Gold standard datasets were used to assist in developing and quality assuring each module's methods, and efficiency was prioritised to enable methods to scale to Big Data.

Results

As of spring 2025, the pre-processing, deterministic, and blocking modules successfully met MVP user requirements, with probabilistic in the final stages. The modules successfully scale to population level datasets, and can be used in succession to clean, block, and link datasets to a person spine with minimal user input and acceptable precision and recall based on gold standard datasets. The modules are flexible, accepting a range of linkage variables and formats, and provide users with the ability to use a default matching strategy or to customise deterministic matchkeys for their own needs; users are also provided with a choice of two probabilistic score thresholds to account for differing quality requirements. Further improvements are ongoing to add additional adaptability and functionality.



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.