

International Journal of Population Data Science

Journal Website: www.ijpds.org



How the RDMF enhances the use of Administrative Data for analysis for the public good

Hannah Goode¹, Elias Kellow¹, Charles Baird¹, Rachael Colquitt¹, Nick Mavron¹, Chris Brooker¹, and David Cobbledick¹

¹The Office for National Statistics, Newport, United Kingdom

Objective

The Reference Data Management Framework (RDMF) connects and unifies linked datasets to create unique indexes for individuals, businesses and locations in the UK. It enables expanded research via secure data matching services, linking data from different data sources, reliable de-identification and other expanded analytical capabilities.

Conclusion

Once a dataset has been indexed, it can link to other indexed datasets through a direct join with the same index id, or through cross index association where all four indexes have been linked together. This allows for scalable and consistent dataset linkage increasing the analytical value of each dataset.

Method

The RDMF uses a series of different approaches to linkage. Fully automated linkage is a direct match on a unique identifier for example a National Insurance Number. Rule based matching applies a set of matching rules which are ordered hierarchically, where the subsequent set of rules is used in the next iteration of linkage. Probabilistic matching calculates the score or probability that two records are a match based on matching agreement weights between variables and calculated through gradient descent. Records are also matched clerically when a computer is unable to decide if two records are a match, a human can decide.

Results

The four indexes created in the RDMF contain one record for each person DI-Demographic Index, business BI-Business Index, location LI-Location Index and industrial and occupational classification CI-Classification Index.

The RDMF is a multi-tool that uses those indexes to allow faster, more consistent linkage of datasets at scale by indexing the data first and then using the index as a common ID to be joined on.

The indexing process, via a combination of automated linkage and matching services, allows for the enrichment of datasets, expansion of their analytical potential and de-identification while retaining analytical use. The RDMF is underpinned by a Quality and Assurance framework made up of a set of policies, activities, processes and outputs in line with the Code of Practice for Statistics.

