

International Journal of Population Data Science

Journal Website: www.ijpds.org



Indexing First Research: understanding the quality of generalised data linkage processes

Esther Lewis¹, Sarah Cummins², Nathan O'Connor¹, and Beth Allen¹

¹Office for National Statistics, London, United Kingdom

²Office for National Statistics, Titchfield, United Kingdom

While data linkage has traditionally used bespoke, customer led methods, an increased demand for linked data has led to the development of indexes representing people, businesses and locations. Data are linked to indexes through generalised methods, and an index ID is appended to each statistical entity in the dataset. Users request relevant data and join them on index IDs. This process is called indexing. This approach has clear benefits, but there is still uncertainty about the quality of indexing and the quality of datasets linked through index IDs. The Indexing First Research programme provides evidence to better understand the quality outputs of these processes and give clear precedents to direct the future application of indexing. It takes existing bespoke linkages and compares them

to an indexing approach on the same project. This provides evidence on the coverage of the indexes, precision and recall of different methods and bias in each linkage method. This paper will set out the aims of the research programme and the progress to date. It will discuss the outcomes of indexing hard to link populations (ie. homeless people and prisoners), of comparing bespoke linkages and linkages via the indexes (ie. births-deaths linkages), and the accuracy of indexing data through generalised methods (ie. nursing data linked to the persons index). This research has implications for the approach taken to data linkage and gives direction for when indexing is appropriate and when bespoke linkage is required.

