

International Journal of Population Data Science

Journal Website: www.ijpds.org



The automated generation of data linkage Information: A reusable, scalable metadata framework for Trusted Research Environments

Mohamed Mhereeg¹, Anna Rawlings¹, Gareth Davies¹, Owen Howell-Wright¹, Lee Hughes¹, Ashley Akbari¹, and Hywel Evans¹

¹Swansea University Medical School, Swansea, United Kingdom

Using the correct linkage keys to join tables is essential when combining data from multiple sources. Although Trusted Research Environments (TREs) provide metadata to support data linkage, to the best of our knowledge, none currently automate the creation of detailed table-level linkage information. Researchers often conduct linkages, and complex linkages exist at many levels. However, this manual approach may misuse available linkages. As the number of data sources and datasets increases in complexity, ensuring correct linkages within and across data sources becomes difficult, reducing efficiency and reproducibility. The framework provides an end-to-end solution for data in the Secure Anonymised Information Linkage (SAIL) Databank by generating a full linkage matrix and documentation directly from common metadata fields (e.g., table_name, column_name, datatype) using configurable pattern-matching rules (e.g., column_name ending

in '_E' denotes an encrypted ID field). The algorithm then compares shared keys across all tables to identify every combination of internal and external linkages, producing linkage matrix with structured natural-language statements for users, which can subsequently be used to generate linkage diagrams. The framework has been validated to ensure accuracy, including verifying that all linkage keys and information are detected. To assess performance, the framework calculates accuracy, precision, and recall by comparing results against a manually generated standard reference, which is validated by the SAIL User-and-Data-Support-Services (UDSS) team. Our approach offers major benefits for TREs and researchers by standardising scalable and reusable linkage documentation. We improved the consistency of linkage metadata production for complex, multi-table datasets within SAIL.

