

International Journal of Population Data Science

Journal Website: www.ijpds.org



SeRP Linkage Centre: A Scalable, Automated Architecture for Record Linkage

Roshan Toby¹, Michael Edwards¹, Anahita Zamani¹, Sean Scully¹, and Simon Thompson¹

¹SeRP, Swansea, United Kingdom; Swansea University, Swansea, United Kingdom

Modern population-scale record linkage and entity resolution requires technical infrastructures that support automation, reproducibility, and secure data integration. Here we present SeRP Linkage Centre, a modular and scalable end-to-end architecture to meet these requirements, enabling standardised linkage workflows and reporting within secure, federated environments. The workflow leverages a structured project directory, designed to standardise inputs, configurations, and outputs across linkage tasks. DBT (Data Build Tool) is used for data cleaning and transformation, providing version-controlled, auditable SQL models that prepare datasets for linkage. The linkage engine is implemented in Splink using a duckdb backend, orchestrated through Airflow DAGs that automate execution, quality checks, and dependency handling. All tasks are containerised and executed on Kubernetes, enabling horizontal scaling and reproducibility. Linkage jobs are triggered using JSON payloads via a REST

API, allowing integration with user interfaces. To support transparency, linkage reports are produced automatically using ReportLab, including model descriptions, summary metrics and match quality indicators, and workflow audit trails from source data to linkage table. The architecture enables high-throughput, repeatable linkage execution with reduced manual overhead. Standardised project structures and containerised infrastructure improve reproducibility and portability across secure environments, whilst automated DBT transformations and reporting enhance traceability and quality assurance. The SeRP Linkage Centre tech stack demonstrates how a modular, automated, and containerised infrastructure can support population-level linkage at scale. Future development includes expanding the range of quality measures and insights, as well as evolving the platform towards a lakehouse architecture to unify storage and metadata governance.

