

International Journal of Population Data Science

Journal Website: www.ijpds.org



Operationalisation of Data Linkage Through Abstracted Workflows

Michael Edwards^{1,2}, Anahita Zamani^{1,2}, and Simon Thompson^{1,2}

¹SeRP, Swansea, United Kingdom

²Swansea University, Swansea, United Kingdom

An abstracted multi-phase, multi-model linkage strategy is presented via an operational workflow applied within a real-world multi-dataset environment, with population-scale datasets spanning multiple domains. The approach provides a consistent, repeatable playbook for data linkage, leveraging the pipeline in producing quality linked data for research in a robust and scalable manner.

At the abstract level, we define building blocks to produce linked cohorts from component models, merging identified sub-graphs into larger cohorts in a systematic pipeline from data acquisition to cohort provisioning. First, each dataset is intra-linked and quality checked, leveraging dataset-specific information to produce high quality within-dataset links. Next, inter-dataset links are produced, utilising common identifiers between sets. Both intra- and inter-set linking phases make use of deterministic and probabilistic linkage methods to produce a comprehensive set of edges. In the final phase, Intra- and inter-set edges are integrated into a single multi-set linked cohort before versioning and release.

Producing a larger cohort from dataset-specific sub-graphs allowed the exploitation of focused models at each stage of the pipeline to improve quality, giving finer control and inspection of potential assumptions and biases during the linkage process. One example was in linkage of a population-scale survey, in which a logical negation excluded possible false links of individuals within a household, who commonly share data features. The approach also allows for a modular, controlled nature to revisions of linkages in improving quality, with modifications being confined to within specific phases of the workflow, with sub-graphs versioned and revised in isolation of the wider linkage. The multi-phase pipeline produces relatively more complex linkages on the whole, however the proposed workflow allows for a systematic approach which supports delivery.

From an operational standpoint, this approach improves linkage turnaround by standardising workflows for linkage operators, abstracts to a wide range of real-world linkage applications, and provides granular control over a cohort's construction at the expense of producing a more complex composition of links through the variety of edge types produced.



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.