International Journal of Population Data Science

Journal Website: www.ijpds.org

A comparison of methodologies: sibling identification using a relational versus a graph-based approach

Farrow, J^{1,2} and Adams, S²

¹Farrow Norris

²SANT DataLink

Introduction

The detection of siblings is an important pre-requisite for many research problems, yet is traditionally difficult or time-consuming owing to the way linked data is conventionally stored. We compare the methodology and results for an identification of siblings by SANT DataLink using a legacy relational approach and a graph-based approach.

Objectives and Approach

An existing project involving the identification of sibling clusters using relational techniques was replicated using the same core data in SANT DataLink's Next Generation Linkage Management System (NGLMS) which is graph-based. Data is stored in the NGLMS using richer relationships than just 'is W% similar to' or 'is part of group N'. Birth data was separated into children and parents and explicit child/parent relationships recorded. When coupled with electoral roll data using probabilistic and deterministic linkage, sibling structure can be identified by performing network traversal via parents, e.g.

`crecord—[MOTHER]—>mrecord—[IS_SIMILAR_TO]—>mrecord*←[MOTHER]—crecord*`

i.e. find the mother, find that mother's cluster, find all children related to that cluster. The resulting records are siblings.

Results

A graph-based approach enabled the methodology of 'finding siblings' to be more clearly described and communicated by mirroring genealogical structures natively within the data. Comparable results were achieved in a shorter time with less manual effort using the same underlying data. Generating sibling clusters from the graph-based data required less manual intervention and review as explicit PARENT/CHILD relationships were stored in the data and able to be quickly traversed to assemble familial units and thus siblings. A focus on automated linkage quality rather than manual review

was facilitated by the approach. Anomalous structures requiring detailed review, such as multiple fathers and mothers of a single child, were trivially identifiable and so manual effort was focussed on actions where a higher return in terms of end-product quality could be achieved.

Conclusion/Implications

Storing data in a representation which more closely resembles the underlying real-world situation allows greater fidelity with respect to data modeling. This in turn, enables the asking of richer questions and makes the answering of such questions much easier.