Evaluating Data Linkage: Creating longitudinal synthetic data to provide 'gold-standard' linked data sets for comprehensive linkage evaluation

Tom Dalton, Graham Kirby, Alan Dearle, Özgür Akgün

School of Computer Science, University of St Andrews, UK
{tsd4, graham.kirby, alan.dearle, ozgur.akgun}@st-andrews.ac.uk

Given that a data set needs to be linked, the true links are by definition unknown. The success of a linkage approach is thus difficult to evaluate. Small hand-linked data sets may be used as a 'gold-standard' for evaluating a linkage approach. However, errors in hand-linkage, and the limited size and availability of such data sets, do not allow for comprehensive evaluation.

Using synthetic data to evaluate linkage algorithms has the advantage that all the true links are known. In the domain of population reconstruction, the ability to synthesise populations on demand, with varying characteristics, allows a linkage approach to be evaluated across a wide range of data sets. Characteristics, for example, include family size, infant mortality, and geographic mobility. The effects of deliberately-introduced transcription errors can also be investigated.

This talk focuses on the creation of synthetic longitudinal data sets, for the assessment of linkage quality achieved by automatic linkage algorithms in the domain of population reconstruction. It presents a micro-simulation model for generating such synthetic populations, taking as input a set of desired statistical properties. It then outlines how these desired properties are verified in the generated populations, and the intended approach to using generated populations to evaluate linkage algorithms. We envisage a sequence of experiments where a set of populations are generated to consider how linkage quality varies across different populations: with the same characteristics, with differing characteristics, and with differing types and levels of corruption. The performance of an approach at scale is also considered.

This work was presented at the following workshop.

Venue	The Workshop for the Systematic Linking of Historical Records
Location	University of Guelph, Canada
Date	May 11-13, 2017
URL	http://recordlink.org