Learning From Past Links: Understanding the Limits of Linkage Quality

Özgür Akgün[1], Alan Dearle[1], Eilidh Garrett[2], Graham Kirby[1]

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

The Digitising Scotland project aims to link 25 million vital event records from 1850s to 1970s. We aim to develop automatic approaches to probabilistic, similarity based record linkage. Linkage quality depends on the choices of keys and similarity measures. However, until now the effect of these choices has been unclear. We study the theoretical limits of automated linkage by performing a post-linkage analysis on two datasets, one from the Isle of Skye and one from Kilmarnock, previously linked by historical demographers. In these datasets, individuals appear on multiple certificates. The linkage problem involves unifying these occurrences e.g. between births and deaths, known as Entity Resolution. This requires the choice of particular keys, a similarity measure and a threshold signalling equivalence. We calculate linkage quality metrics–precision, recall, and f-measure–for 4 different key combinations, different similarity measures, and a range of threshold values. We present the distribution of similarity values for links and non-links for each configuration and data-set. From these results, we hope to understand the limits of automated probabilistic record linkage. We will use this understanding to inform our approach to the linkage of new unlinked datasets such as the Digitising Scotland dataset. We would welcome the opportunity to apply this approach to other linked demographic datasets.

This work was presented at the following workshop.

Venue	BSPS 2017 - British Society of Population Studies Annual Conference
Location	University of Liverpool
Date	6 - 8 September 2017