Linking Scottish vital events records back through time to produce the 'Understanding Scotland's People Study'

Chris Dibben, Zhiqiang Feng, Zengyi Huang, Graham Kirby & Lee Williamson

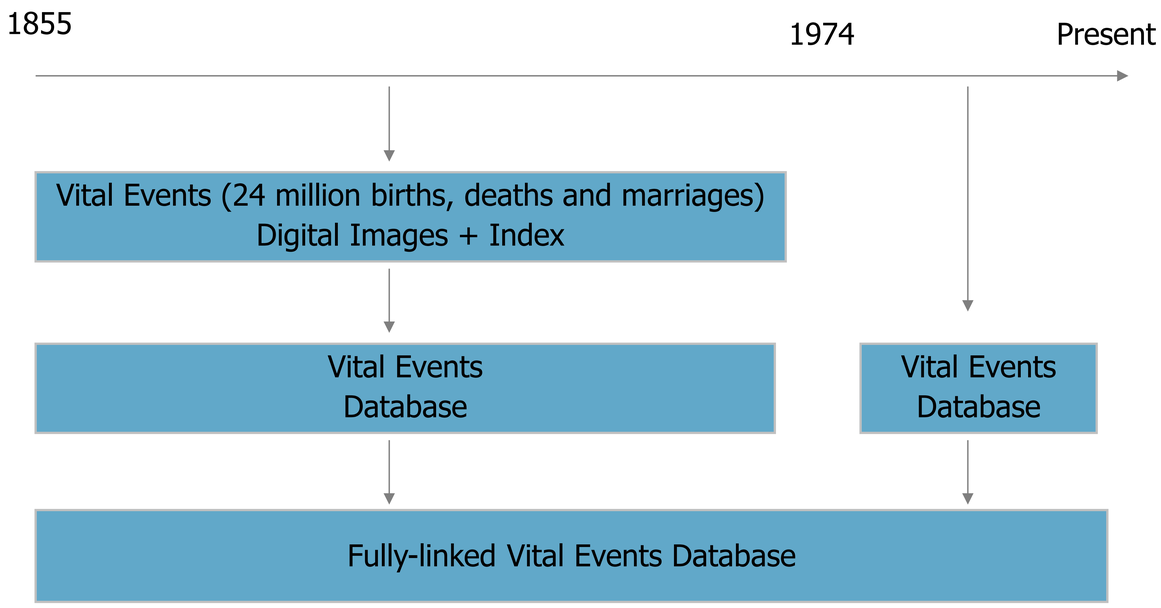
University of St Andrews

Introduction

The '*Understanding Scotland's People Study*' will involve the construction of a multidisciplinary research database containing information on some 18 million individuals. It will be a multi-generational, lifecourse dataset, showing familial relationships and containing health (cause of death), fertility, socio-economic (occupation) and geographical data for the whole of Scotland from 1855 to the present day.

Project outline

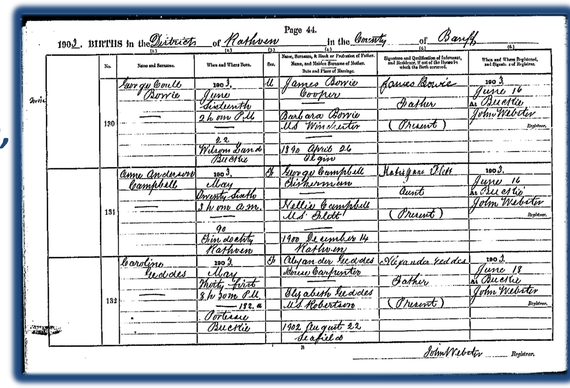
The work will involve the [**stage 1**] transcription of some 24 million vital events record images (births, deaths and marriages) for the whole of Scotland back to 1855 and [**stage 2**] the linking of these records for individuals and then between parents and their children to construct a lifecourse and pedigree/genealogy dataset for most of the population who have ever lived in Scotland between 1855 and the present day and finally [**stage 3**] the coding of occupation, causes of death and place of residence to be research usable.



Project progress

Stage 1

Digitising Scotland: the DIGROS (Digital-Imaging of the Genealogical Records of Scotland's People) electronic index enhancement, the vital events birth, death and marriage registers from 1855-1973



- Funding is in place for the transcription of the 24 million vital events record images to proceed, at present the work is being tendered for.
- The transcription should be completed in 2014

Stage 2

Linking certificates from 1855 into pedigrees/family trees

- In collaboration with computer scientists we are exploring novel approaches to processing and storing large-scale genealogical data and to representing and reasoning about the inherent uncertainties in such data (for example, graph theory).
- At present we are considering creating synthetic family structure data in order to develop the knowledge base and rule engine necessary for achieving the linkage.

Stage 3

Coding of occupation and causes of death

- We are working with historians and computer scientists to address issues surrounding the standardisation of causes of death and occupations with the aim of classifying the standard into an internationally comparative scheme (ie ICD-10 and HISCO).
- Currently investigating natural language processing (NLP) and machine learning tools.
- Progress suggests although NLP was initially promising, the Apache's Mahout toolkit provides a number of tools and algorithms suitable for classifying the historical cause of death strings. We are experimenting with different machine learning algorithms in order to find which approach gives the most confident results and the best accuracy.

Further data linkage

- The '*Understanding Scotland's People Study*' dataset will be prepared for linkage to the already highly developed Scottish health informatics systems and will therefore enhance contemporary Scottish and UK health datasets, health informatics systems, longitudinal datasets and genetic studies as well as potentially allowing other historical datasets to be linked to these contemporary records.