Reconstructing Genealogical Relationships from Public Registers

Graham Kirby, Alan Dearle, Masih Hajiaraab Derkani, Tom Dalton, Konstantinos Daras
University of St Andrews

Chris Dibben, Lee Williamson, Zengyi Huang, Zhiqiang Feng
University of Edinburgh

Eilidh Garrett, Alice Reid
University of Cambridge
Background

• Digitising Scotland project
  • transcribe vital events records 1855-1973
    • births
    • deaths
    • marriages
  • codify occupations, causes of death, locations
  • link records to form family tree(s)
• www.lscs.ac.uk/projects/digitising-scotland/
Why?

• To produce resource for further research
  • patterns of migration within Scotland
  • social mobility
  • impact of developing welfare state
  • impact of industrialisation
  • distributions of occupations
  • effects of occupation on health
  • long-term impacts of serious viruses
  • genetic influences on disease
  • …
Scottish Public Records

- Recorded nationally since 1855
- Birth records: 14 million
- Marriage records: 4 million
- Death records: 11 million
- About 18 million individuals
<table>
<thead>
<tr>
<th>No.</th>
<th>Name and Surname</th>
<th>When and Where Born</th>
<th>Sex</th>
<th>Name, Surname, &amp; Rank or Profession of Father.</th>
<th>When and Where Registered, and Signature of Registrar.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1360</td>
<td>Charles Rennie</td>
<td>1868</td>
<td></td>
<td>William McIntosh</td>
<td>1868 at Glasgow</td>
</tr>
<tr>
<td></td>
<td></td>
<td>in Glasgow</td>
<td></td>
<td>Police Clerk</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1361</td>
<td>Agnes Caldwell</td>
<td>1862</td>
<td></td>
<td>William Caldwell</td>
<td>1862 at Glasgow</td>
</tr>
<tr>
<td></td>
<td></td>
<td>in May</td>
<td></td>
<td>Mrs. M. St. Blyth</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1362</td>
<td>Electoral</td>
<td>1868</td>
<td></td>
<td>Alexander Jackson</td>
<td>1868 at Glasgow</td>
</tr>
<tr>
<td></td>
<td></td>
<td>in June</td>
<td></td>
<td>Railway Clerk</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

www.scotlandspeople.gov.uk/content/images/famousscots/birthCharlesRMcIntosh1868.tif
<table>
<thead>
<tr>
<th>No.</th>
<th>Name and Surname</th>
<th>When and Where Born</th>
<th>Sex</th>
<th>Name, Surname, &amp; Rank or Profession of Father</th>
<th>Name, and Maiden Surname of Mother</th>
<th>Date and Place of Marriage</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Charles Rennie</td>
<td>1868</td>
<td>M</td>
<td>William McIntosh</td>
<td>Police Clerk</td>
<td></td>
</tr>
<tr>
<td></td>
<td>McIntosh</td>
<td>June</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1360</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Rose Ann</td>
<td>1868</td>
<td>F</td>
<td>William Caldwell</td>
<td>Snow Moulder</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Caldwell</td>
<td>May</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Name</td>
<td>Place of Birth</td>
<td>Occupation</td>
<td>Father's Name</td>
<td>Mother's Name</td>
<td>Residence</td>
<td>Date of Marriage</td>
</tr>
<tr>
<td>-------</td>
<td>----------------</td>
<td>------------</td>
<td>---------------</td>
<td>--------------</td>
<td>------------</td>
<td>-----------------</td>
</tr>
<tr>
<td>John Doe</td>
<td>Edinburgh</td>
<td>Lawyer</td>
<td>Robert Smith</td>
<td>Jane Doe</td>
<td>123 Main St.</td>
<td>1850</td>
</tr>
<tr>
<td>Jane Smith</td>
<td>London</td>
<td>Teacher</td>
<td>Edward Jones</td>
<td>Sarah Smith</td>
<td>456 Elm St.</td>
<td>1850</td>
</tr>
</tbody>
</table>

Notes:
- Marriage performed according to the Church of Scotland.
- Both parties were of sound mind and free will.
- Witnesses: William Brown and Mary Jackson.

1880. MARRIAGES IN THE GHILDRED OR STIRLING.
<table>
<thead>
<tr>
<th>No.</th>
<th>Name and Surname</th>
<th>When and Where Died</th>
<th>Sex.</th>
<th>Age</th>
<th>House of Death, Duration of Disease, and Medical Attended</th>
<th>Signature &amp; Qualification of Informant and Registrar, If out of the House in which the Death Occurred</th>
</tr>
</thead>
<tbody>
<tr>
<td>130</td>
<td>George ^McIntosh^</td>
<td>February 1880</td>
<td>M</td>
<td>83</td>
<td>Geo. McIntosh, Compl. of Fall</td>
<td>George G. Mcintosh</td>
</tr>
<tr>
<td>130</td>
<td>Henry ^McIntosh^</td>
<td>February 1880</td>
<td>M</td>
<td>83</td>
<td>Geo. McIntosh, Compl. of Fall</td>
<td>George G. Mcintosh</td>
</tr>
<tr>
<td>130</td>
<td>James ^McIntosh^</td>
<td>February 1880</td>
<td>M</td>
<td>83</td>
<td>Geo. McIntosh, Compl. of Fall</td>
<td>George G. Mcintosh</td>
</tr>
<tr>
<td>130</td>
<td>John ^McIntosh^</td>
<td>February 1880</td>
<td>M</td>
<td>83</td>
<td>Geo. McIntosh, Compl. of Fall</td>
<td>George G. Mcintosh</td>
</tr>
<tr>
<td>130</td>
<td>Joseph ^McIntosh^</td>
<td>February 1880</td>
<td>M</td>
<td>83</td>
<td>Geo. McIntosh, Compl. of Fall</td>
<td>George G. Mcintosh</td>
</tr>
<tr>
<td>130</td>
<td>Robert ^McIntosh^</td>
<td>February 1880</td>
<td>M</td>
<td>83</td>
<td>Geo. McIntosh, Compl. of Fall</td>
<td>George G. Mcintosh</td>
</tr>
<tr>
<td>130</td>
<td>William ^McIntosh^</td>
<td>February 1880</td>
<td>M</td>
<td>83</td>
<td>Geo. McIntosh, Compl. of Fall</td>
<td>George G. Mcintosh</td>
</tr>
<tr>
<td>130</td>
<td>Mary ^McIntosh^</td>
<td>February 1880</td>
<td>F</td>
<td>83</td>
<td>Geo. McIntosh, Compl. of Fall</td>
<td>George G. Mcintosh</td>
</tr>
<tr>
<td>130</td>
<td>Isabella ^McIntosh</td>
<td>February 1880</td>
<td>F</td>
<td>83</td>
<td>Geo. McIntosh, Compl. of Fall</td>
<td>George G. Mcintosh</td>
</tr>
</tbody>
</table>

**Note:** The table contains a list of deaths with details such as the name, sex, age, cause of death, and the signature of the registrar. The data is from the DEATHS in the District of St. Mary in the Borough of Dumbarton for the year 1880.
Problem: Codification
Problem: Codification

• Classify causes of death to ICD10 (about 1500 codes)
Problem: Codification

- Classify causes of death to ICD10 (about 1500 codes)
  - “cerebral haemorrhage”
Problem: Codification

• Classify causes of death to ICD10 (about 1500 codes)
  • “cerebral haemorrhage”
  • “cancer of colon”
Problem: Codification

- Classify causes of death to ICD10 (about 1500 codes)
  - “cerebral haemorrhage”
  - “cancer of colon”
  - “heart failure”
Problem: Codification

- Classify causes of death to ICD10 (about 1500 codes)
  - “cerebral haemorrhage”
  - “cancer of colon”
  - “heart failure”
  - “effusion on the brain caused by excessive drinking and exposure to cold”
Problem: Codification

• Classify causes of death to ICD10 (about 1500 codes)
  • “cerebral haemorrhage”
  • “cancer of colon”
  • “heart failure”
  • “effusion on the brain caused by excessive drinking and exposure to cold”
  • “drowned by a smack on boat having been culpably negligently or recklessly navigated or steered by Angus McPherson boatman, Portree”
Problem: Codification

- Classify causes of death to ICD10 (about 1500 codes)
  - “cerebral haemorrhage”
  - “cancer of colon”
  - “heart failure”
  - “effusion on the brain caused by excessive drinking and exposure to cold”
  - “drowned by a smack on boat having been culpably negligently or recklessly navigated or steered by Angus McPherson boatman, Portree”
  - “drowning possibly through his falling asleep or getting insensible through liquor on the sea shore at low water and being afterwards covered by the sea on the tide flowing”
Problem: Tree Reconstruction

- From the set of certificates, need to work out:
  - how many people?
  - what were the relationships?
Problem: Tree Reconstruction

- Match up information on certificates
  - may be ambiguous
- information may be missing
  - mistakes
  - immigration and emigration
Problem: Tree Reconstruction

- Birth Certificate 3826756
  - Jane Smith, born 1/3/1889

- Marriage Certificate 573734
  - Janet Smyth, born 11/3/1889

- How many people?

- Also consider other information e.g. place of birth, occupation, relationships to other people
Problem: Scale
Problem: Scale

• Even assuming you can decide whether a given pair matches, there are a lot of possible pairs to match…
Problem: Scale

• Even assuming you can decide whether a given pair matches, there are a lot of possible pairs to match…

• 3 people mentioned on a birth certificate
Problem: Scale

• Even assuming you can decide whether a given pair matches, there are a lot of possible pairs to match…
  
  • 3 people mentioned on a birth certificate
  
  • 6 on a marriage certificate
Problem: Scale

- Even assuming you can decide whether a given pair matches, there are a lot of possible pairs to match…
  - 3 people mentioned on a birth certificate
  - 6 on a marriage certificate
- $3 \times 14 \text{ million} \times 6 \times 4 \text{ million} = 10^{15}$
Problem: Scale

• Even assuming you can decide whether a given pair matches, there are a lot of possible pairs to match…
  • 3 people mentioned on a birth certificate
  • 6 on a marriage certificate
  • $3 \times 14\text{ million} \times 6 \times 4\text{ million} = 10^{15}$
  • If we could check a pair in a microsecond this would take 32 years
Problem: Scale

• Even assuming you can decide whether a given pair matches, there are a lot of possible pairs to match…
  • 3 people mentioned on a birth certificate
  • 6 on a marriage certificate
  • $3 \times 14 \text{ million} \times 6 \times 4 \text{ million} = 10^{15}$
  • If we could check a pair in a microsecond this would take 32 years
  • Also need to check births against births & deaths, deaths against deaths, births, marriages, …
Problem: Evaluation
Problem: Evaluation

• How do you know whether the reconstructed tree is correct?
Problem: Evaluation

• How do you know whether the reconstructed tree is correct?
• In general, you can’t: very little ground truth
  • the whole point of doing this automatically is that we don’t have time for experts to do it
• but we can get experts to look at some parts
Problem: Evaluation

• How do you know whether the reconstructed tree is correct?

• In general, you can’t: very little ground truth

• the whole point of doing this automatically is that we don’t have time for experts to do it

• but we can get experts to look at some parts

• For intended research purposes, doesn’t matter as long as it’s reasonably close
Problem: Evaluation

• How do you know whether the reconstructed tree is correct?

• In general, you can’t: very little ground truth
  • the whole point of doing this automatically is that we don’t have time for experts to do it
  • but we can get experts to look at some parts

• For intended research purposes, doesn’t matter as long as it’s reasonably close

• Also possible to evaluate linkage techniques with synthetic data
Problem: Understanding Results
Problem: Understanding Results

- Sometimes we won’t be able to make good automatic matching decisions
- Need to record this uncertainty and let researchers know about it
Problem: Understanding Results

• Sometimes we won’t be able to make good automatic matching decisions

• need to record this uncertainty and let researchers know about it

• How to present this in a way that’s easy to understand?

• multiple possible family trees for a given person
Summary

• **Aim**
  
  • produce linked data set containing all people born, married or died in Scotland 1855-1973

• **Problems**
  
  • transcription of data
  
  • classification of occupations and causes of death

• **Research issues**
  
  • techniques for linking accurately and efficiently
  
  • understanding the consequences of uncertainty