Record Linkage in Medieval and Early Modern Documents

ChartEx and Traces Through Time project
Arno Knobbe, Universiteit Leiden
Two Digital Humanities Projects

ChartEx, 2012-2013 (*Digging into Data* challenge)
- U. Leiden (Data Mining)
- U. Brighton (NLP)

Traces Through Time, 2014-2015 (*Big Data, AHRC*)
- The National Archives, U. London
- U. Leiden
- U. Brighton
ChartEx, the Charter Excavator

- Over 20k charters from various regions
  - England, mostly York
  - France, Cluny
- 5 collections of charters
  - modern translations to English
  - Latin
- Middle ages
  - 900-1400
408. Grant by Thomas son of Josce goldsmith and citizen of York to his younger son Jeremy of half his land lying in length from Petergate at the churchyard of St. Peter to houses of the prebend of Ampleford and in breadth from Steyngate to land which mag. Simon de Evesham inhabited; paying Thomas and his heirs 1d. or [a pair of] white gloves worth 1 d. at Christmas. Warranty. Seal.
January 1252 [1252/3]

SOURCE: VC 3/Vi 326 (161 mm x 137 mm)
408. Grant by Thomas son of Josce goldsmith and citizen of York to his younger son Jeremy of half his land lying in length from Petergate at the churchyard of St. Peter to houses of the prebend of Ampleford and in breadth from Steyngate to land which mag. Simon de Evesham inhabited; paying Thomas and his heirs 1d. or [a pair of] white gloves worth 1 d. at Christmas. Warranty. Seal.


January 1252 [1252/3]
408. Grant by Thomas son of Josce goldsmith and citizen of York to his younger son Jeremy of half his land lying in length from Petergate at the churchyard of St. Peter to houses of the prebend of Ampleford and in breadth from Steyngate to land which mag. Simon de Evesham inhabited; paying Thomas and his heirs 1d. or [a pair of] white gloves worth 1 d. at Christmas. Warranty. Seal.


January 1252 [1252/3]
Record Linkage Challenges

- Fairly unstructured data, natural language

- No concept of last names (Middle Ages)
  - *Thomas*
  - Place names, *William de Gerfordby*
  - Occupation, *Robert le Spicer*

- No street numbers (nor GPS)
  - *land with buildings in Petergate, lying in length and in width between the land once of Nicholas de Bugetorp and the land once of Martin Grandeth*

- Spelling variations
  - *John de Eskrick, John de Eskryck, John de Escryck*
408. Grant by Thomas son of Josce goldsmith and citizen of York to his younger son Jeremy of half his land lying in length from Petergate at the churchyard of St. Peter to houses of the prebend of Ampleford and in breadth from Steyngate to land which mag. Simon de Evesham inhabited; paying Thomas and his heirs 1d. or [a pair of] white gloves worth 1d. at Christmas. Warranty.
Several NLP layers

**Syntax layer:** build (local) syntactic structure to identify basic constituents of the sentence.
**Phrasal layer:** use part-of-speech tags to build lexical items into local syntactic/semantic structures
Record Linkage

- Consider pairs of occurrences
  - From two documents, or within one document
  - Are these the same person?
- List matching evidence
  - First name, last name, occupation...
- The more evidence, the higher the confidence
- Assign a confidence value

... John Smith ...  ... Mr. John Smith ...
Matching Relational Information
Thomas son of Josce matched
Probabilistic Record Linkage

- Consider pairs of occurrences
  - Are these the same person?
  - Assign a confidence value
- Probabilistic: judge available evidence
  - Every piece of evidence has statistics
  - How common is ‘John’, how common ‘Smith’, etc.?
  - Statistics on
    - First names, last names, titles, roles, occupations, provenance (=size of city)
- Compute confidence of joint evidence
  - Assumes independence of evidence
Medieval First Names

1528 cases
Early Modern Title Statistics

1664 cases
Medieval example

Osbert 0.0021
  Father: Ralph 0.0342
  Grandfather: Hugh 0.028

Matched with:
- Osbert, son of Ralph, grandson of Hugh
  - conf = 0.508
  - roughly two such people at the time

- Osbert de Somere, son of Ralph
  - conf = 0.0142
  - many people are actually ‘Osbert, son of Ralph’ (approx. 70)
Challenges

- Getting reliable statistics
  - source data gives clue about frequency
  - but biased: an entity (VIP) may have multiple occurrences
  - iterative method
    - assume statistics from source text
    - link best occurrences
    - update statistics

- Independence assumption?
  - First names vs. last names
  - Clustering over subpopulations (English, Scottish, Welsh, Jewish, …)
  - Last names vs. provenance, last names vs. titles, etc.
Conclusions

- Natural Language Processing
  - moderately successful
  - work in progress

- Record Linkage doable
- Informally tested on small data
- Many more opportunities when done ‘big’

- Opportunities for Digital Humanities