Entity Resolution in Unstructured Data

and applications in the analysis of historical documents

Benjamin van der Burgh

March 15th 2016

Supervisors: Dr. Arno Knobbe
Dr. Siegfried Nijssen

Overview

1. Goals of the Traces Through Time project
2. Format problem description
3. Record extraction
4. Comparison of record fields
5. Candidate pair classification
6. Maximally k-informative itemsets
7. Experiments
8. Conclusions and future work

Universiteit Leiden

Traces Through Time (1) – Context

- The National Archives stores millions of documents.
- Many documents have been converted to a digital format.
  - Automatic: Optical Character Recognition (OCR).
  - Manual: transcribed by hand.
- Connecting pieces of information regarding people is mostly done manually.
- Automating this process allows for studying people in all layers of society, not just the aristocracy.

Universiteit Leiden

> *No matter what he does, every person on earth plays a central role in the history of the world. And normally he doesn't know it.*
>
> *Paulo Coelho (The Alchemist)*

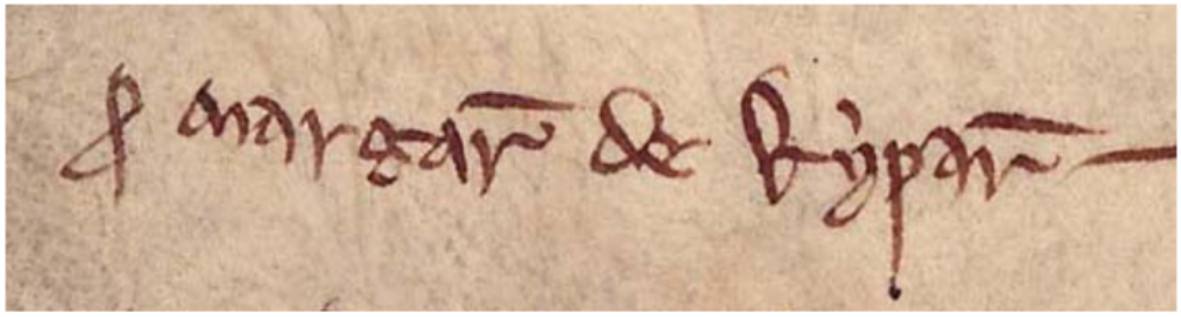
Universiteit Leiden

Traces Through Time (2) – Goals

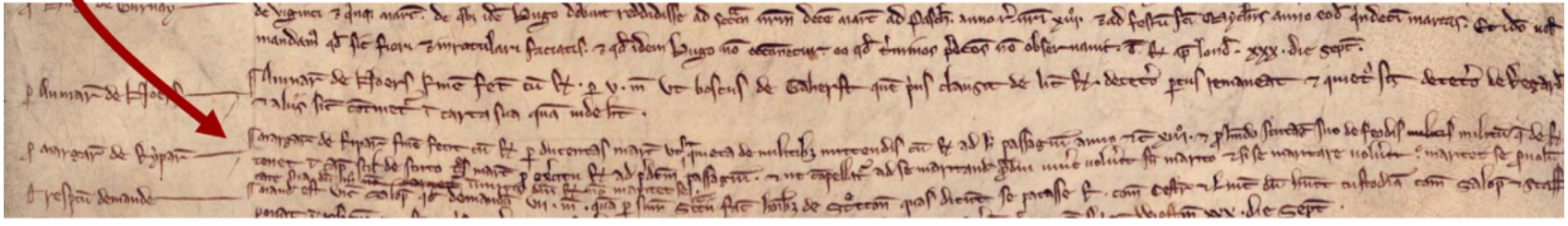
- Develop a methodology to identify and trace individuals across large and diverse historical datasets.
- Look particularly at ‘fuzzy’ data
  - Aliases: Will, William
  - Incomplete data: John (only a name)
  - Spelling variations: Owen, Eoghan
  - (OCR) Errors: Wihiam (William)

Universiteit Leiden

Margaret de Redvers



325 [**No date**]. *For Margaret de Redvers*. Margaret de Redvers has made fine with the king by 200 m., so that she is to be quit of sending knights with the king at his passage in the thirteenth, year, and for having her scutage from the knights' fees that she holds of the king in chief, namely 3 m. per shield for the king's army at the aforesaid passage, and so that she shall not be compelled to marry for as long as she wishes to live without a husband, and if she will wish to marry, she is to marry by her will on condition that she does not marry enemies of the king. [1]



Traces Through Time (3) – Collaboration

- The project set out as a collaboration between several institutes:
  - The National Archives
  - Institute of Historical Research
  - Brighton University
  - Leiden University
- Brighton University worked on *Natural Language Processing*.
- Our job was to perform record linkage on the extracted references delivered by Brighton University.

Universiteit Leiden

Problem Definition (1)

Record

A record r is a tuple of m attributes, each having a certain domain, that describes an entity, i.e., $r \in A_1 \times A2 \times \cdots \times A_m$.

- We assume that records are descriptions of people.
- Records are potentially ambiguous: they can describe more than one person.

Universiteit Leiden

Problem Definition (2)

Record Linkage

Given a set $\mathcal{R}$ of records, determine which of these records refer to the same entity.

- Record linkage is a binary classification problem.
- Record pairs are classified as matching or non-matching.
- The set of entities is usually unknown.
- Even with expert knowledge, it is hard to determine the match status of a record pair.

Record Extraction

- Instead of waiting for input from Brighton University, a simple context-free grammar was written in order to extract occurences.
- First names and articles (of, de la, etc.) were used as anchor points in the text.
- Capitalization, punctuation and ordering define the class of surrounding words.

{first name} {article} {capitalized word}

↓

{first name} {article} {last name}

Universiteit Leiden

Record Examples

Concerning the corn of Roger of Hyde. Order to the sheriff of Oxfordshire to make the king's advantage without delay, by the view of law-worthy men, from all of the corn of Roger of Hyde, knight, in Hyde, who is with the Earl Marshal, and to put in gage etc. all those who he will find threshing that corn and intermeddling with the land of the same Roger without warrant, to be before the king at his command to answer for it.

Title	First name	Article	Last name	Role
	Roger	of	Hyde	
sherrif		of	Oxfordshire	
	Roger	of	Hyde	knight
	Roger			

Universiteit Leiden