An Interactive, Web-based Tool for Genealogical Entity Resolution

Julia Efremova [a] Bijan Ranjbar-Sahraei [b] Frans Oliehoek [b]

Toon Calders [c] Karl Tuyls [b]

[a] *Eindhoven University of Technology, Eindhoven, The Netherlands*
[b] *Maastricht University, Maastricht, The Netherlands*
[c] *Université Libre de Bruxelles, Bruxelles, Belgium*

Abstract

We demonstrate an interactive, web-based tool which helps historians to do *Genealogical Entity Resolution*. This work has two main goals: First, it uses Machine Learning (ML) algorithms to assist humanities researchers to perform Genealogical Entity Resolution. Second, it facilitates the generation of benchmark data for computer scientists to improve available ML-based Entity Resolution techniques.

1 Introduction

Consider a person named *Theodorus Werners* born in *Tilburg* on *August 11th, 1861*. He got married to *Maria van der Hagen* in *1888*. *Maria Eugenia Johanna Werners* was their child, born in *Tilburg* in *October 1894*. Two years after child's birth, they bought a house in *Breda*. *Theodorus* died in *Breda* on *September 1st, 1926*. Each of these pieces of information might have been mentioned in a structured document such as Birth, Marriage or Death certificate, or a free text document such as a Notarial Act. However, due to changes in spelling conventions, misspellings, data conversion and data loss, linking the name-references that are associated with the same entities (i.e. Entity Resolution (ER)) is a long standing open challenge [1]. Incorporating experts knowledge can be a solution for challenges of ER in uncertain data [2]. Therefore, there is a mandatory need for interactive visual analytic tools that can assist experts with making recommendations to limit their search space and preparing quick visual comparisons. The rest of this paper describes how we've developed such an interactive visual tool for assisting historical experts.

2 Data Setup and Developed Tool

The genealogical data, used in this project, provided by Brabants Historical Information Center, is comprised of two major sources: The first source is a collection of the Birth, Marriage and Death certificates belonging to North Brabants, a province of the Netherlands, for the period of 1700 to 1920 (in total around 1,900,000 certificates). The second source consists of around 90,000 free text documents, mostly the notarial acts, of the province before 1920.

The developed tool, built using the Django[1] framework, uses various programming tools for storage, exploration and refinement of available data. It benefits from an intelligent searching engine, developed based on the Solr[2] enterprise search platform, with which historians can easily search through the dataset.

[1] https://www.djangoproject.com/
[2] http://lucene.apache.org/solr/

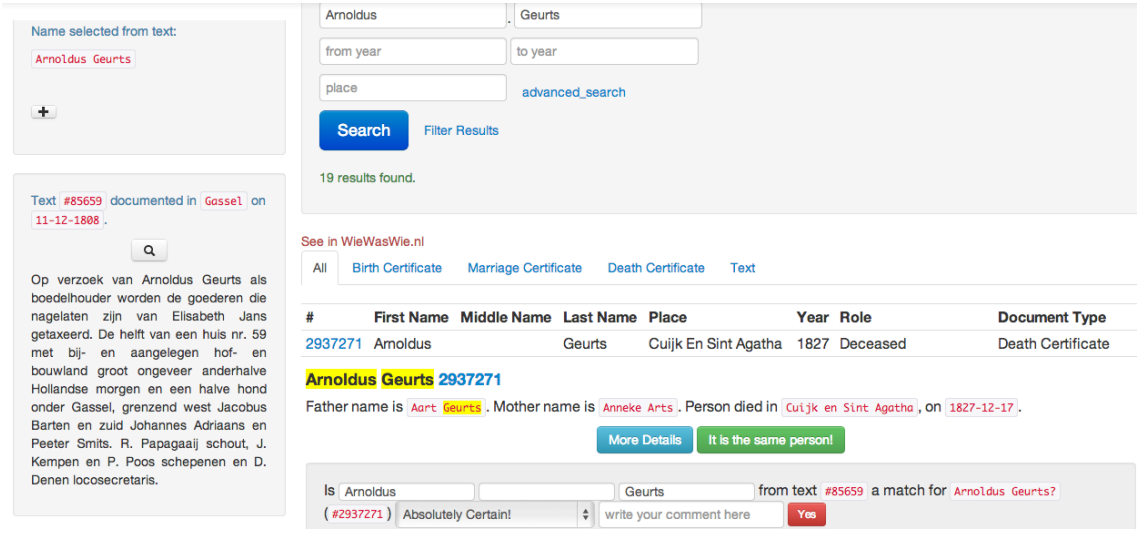

Figure 1: Developed web-based interface: The historian is confirming a match between the name in notarial act (on left) and a name-reference from a Death certificate (on right).

Basically, the required data can be found via person name, location, date and relations. While, the intelligent searching tool lets historians to run complex queries such as *"A person who has married in Breda, Born in Tilburg, and died in 1908"*, as well.

On top of the intelligent searching tool, an easy to use yet powerful **Labeling Tool**, shown in Figure 1, is built which assists historians to link name-references mentioned in a document to name-references in other documents, and consequently expands the available social network (i.e. combinations of sparsely connected pedigrees). Features of the developed labeling tool include automatic name recognition from free text document, suggestion of potential matches, visualization of revealed pedigrees, and detailed comparisons between name references.

In practice, the time required to report a correct match between two name-references varies from a few seconds to probably hours of time. This depends on how similar two references are (e.g., whether places, dates, ages, professions and relatives match or not), and how easy it is to compare those two references. Consequently, the level of confidence in reporting a match varies. Therefore, the actions that historians take (e.g., which keywords they search for and how fast they can recognize a match), and their level of confidence in reporting the match are all stored in the database. As a result, a rich benchmark is generated that includes the list of matches, the level of confidence and list of the actions that historians take before reporting the match. This benchmark data can be used to develop novel ML-based ER techniques and also to measure the accuracy of currently available ER approaches.

3 Demonstration

Demonstration of the developed tool needs a browser and an Internet connection. After introduction of the designed interactive tool, audiences can use the tool and different available features to navigate the available genealogical data, and contribute to the benchmark, while getting familiar with various online dynamic visualizations. The demo will take approximately 20 minutes.

References

[1] Matthew A Jaro. Advances in record-linkage methodology as applied to matching the 1985 census of tampa, florida. *Journal of the American Statistical Association*, 84(406):414–420, 1989.

[2] Mustafa Bilgic, Louis Licamele, Lise Getoor, and Ben Shneiderman. D-dupe: An interactive tool for entity resolution in social networks. In *Visual Analytics Science And Technology, 2006 IEEE Symposium On*, pages 43–50. IEEE, 2006.