Master's thesis in

Applied Data Science

Utilizing linked open place data for entity resolution: a case study on Dutch genealogy at CBG

Author:
S. Engelberts
1422138
s.engelberts@students.uu.nl

Primary supervisor:
prof. dr. A.H.L.M. Pieters
Utrecht University
t.pieters@uu.nl

Secondary supervisor:
S.B. Dirks MSc
Utrecht University
s.b.dirks@uu.nl

External supervisor:
P. Woltjer
Centre for Genealogy
Pieter.Woltjer@cbg.nl

June 29, 2022

Abstract

Entity resolution on genealogical documents is challenging due to spelling errors, alternative name variants, and historic entity changes. Traditional methods attempt to tackle these problems with string similarity methods, which this research proposes to extend by enriching the recorded features with additional place information such as place URIs, coordinates, and country indicators. Based on a case study at the Dutch Centre for Genealogy, this research contributes to extending entity resolution research, optimizing and enriching family history (meta) studies, and investigating which privacy-sensitive passport request documents can be disclosed.

First, linked open data sources are shown to retrieve unique place entities belonging to recorded place names. Second, place, province and country name similarities are calculated as well as coordinate distances within a coordinate reference system that limits the distance distortions for the respective countries. Third, the researched adaptation is shown to result in a significant change in similarity values when a uniform weighting of feature similarities is applied. However, contrary to the hypothesis, the similarity distributions of compared documents that do and do not refer to an equivalent person entity could not be distinguished in a more accurate way. Hence, future studies are proposed that expand on this research by supervised learning of weights and thresholds using validated candidate links from this research.

Keywords: Entity resolution, Linked open data, Genealogy, Geospatial data, Content similarity

Contents

Chapter 1

Introduction

Interest into family history has increased greatly since the twentieth century, giving rise to organisations that collect such information and make this increasingly more easily (digitally) available (CBG, 2016f; Pine, 2021). In the Netherlands, Centre for Genealogy (CBG) aids in family history research by maintaining and digitally publishing genealogical documents (CBG, 2016g). Examples of their collections are personal record cards containing personal information about deceased people, Dutch passport requests by Indonesian-Dutch people after the independence of Indonesia, police records, prayer cards to remember an event for Catholics, World War II documents, et cetera (CBG, 2016a).

1.1 Problem description

There exist some shortcomings with the current state of family history research and the archiving of these collections. Firstly, one has to search for a person or family name in each collection separately (CBG, 2016b). Such a query may return many documents of people with (almost) the same name, from which one has to select their correct ancestor based on the available information. This procedure then has to be repeated for each ancestor to be able to create a full family tree with interesting records about their lives.

Secondly, to be legally allowed to disclose privacy-sensitive documents to descendants, which genealogical documents with personal information are, recorded people should have passed away already according to the General Data Protection Regulation (GDPR) of the European Union and its Dutch variant Algemene Verordening gegevensbescherming (AVG) (GDPR.eu, 2022; Schermer et al., 2018). Additionally, the Dutch Archiefwet specifies that such civil status records have the following legal disclosure periods: 100 years after birth, 75 years after marriage, and 50 years after death (JenV, 1998; van Koutrik & Welings, 2019). Due to these laws it is important to verify with other documents, such as personal record cards at CBG, if a

person is deceased or if these periods have passed.

Thirdly, due to the vast amount of genealogical documents it is infeasible for domain experts to manually link all collections to each other. This is already complicated for one family, so relatedness research to track down heirs for notaries is currently relatively expensive (CBG, 2016c). Consequently, research is also limited to specific families instead of performing meta studies about all families in a region. Examples of such meta studies are: shifts in ages someone gets married or gets children (Rahmani et al., 2014; Rahmani et al., 2016), family migration (Cuijuan et al., 2018), studying lifespan and other factors in relation to socio-economic conditions, and (linguistic) regional or temporal differences in person names.

1.2 Objective

To partially solve the discussed problems from Section 1.1, this research attempts to design a data processing pipeline that can interlink genealogical collections, while utilizing additional information about places that is not mentioned in the documents themselves. The ultimate goal is that one ancestor query returns all results belonging to this specific person from multiple collections at once, as well as their relationships to other people, locations, and occupations. In contrast, at the moment each document is identified based on a person's name, which is not a unique identifier to know if two documents that mention the same name also refer to the same real-world person. Similarly, it is unknown if, for example, referenced names of places or relatives refer to the same real-world entities[1].

When better can be distinguished which documents (do not) refer to equivalent entities, less mistakes are made with linking documents about a person entity. This is especially important for privacy sensitive documents such that none get freely disclosed if the referenced people are still alive. However, such personal information is very valuable for descendants, so any document that can correctly be made available is beneficial.

Additionally, for scientists it is important to continuously improve on methods that can identify unique entities from documents such that knowledge about the world can get accurately enriched and interconnected. Such an approach can also be applied within other domains than genealogy, and to connect documents of relatives who are mentioned on documents.

1.3 Challenges and technical domain

Retrieving additional information about places and connecting genealogical documents that refer to the same entities is a challenging task. This due

[1]Entities are unique things with features that describe them, such as people, places, and products (Goyal, 2021).

to spelling errors, not all documents containing the same (meta) data, a large number of documents (Goyal, 2021), temporal evolution of names of people and places with respect to spelling variations and alternative names (den Engelse, 2015b; Ehrmann et al., 2021), different entities having the same name, and evolution of entities as a whole like places that fused with nearby places (den Engelse, 2015b; Sehgal et al., 2006).

To tackle these challenges and the objective of Section 1.2, this research focuses on a combination of entity resolution (ER) and linked open data (LOD). ER is a procedure that attempts to link documents that refer to the same entities on the basis of recorded features (Goyal, 2021). Further, LOD structures information about unique entities in a standardized way, such that it is unambiguous what relationships exist between entities and their features, and stated explicitly which unique entity is referred to (Berners-Lee, 2009; Blaney, 2021). LOD can thus enrich the recorded information about (place) entities, and help verify if the same or a similar one is referenced in compared documents.

1.4 Research questions and hypothesis

Now the problem, objective, challenges, and technical domain are introduced, the research question can be defined as:

> *How does enrichment of genealogical data with place information influence entity resolution on personal documents?*

To answer this research question, the following subquestions are addressed:

1. How can external information about places be retrieved?
2. Which place features and corresponding similarity metrics are shown to perform well for identifying equivalent places?

This research hypothesizes that enrichment of genealogical data with place information increases the certainty with which (non-)equivalent place entities can be identified, and tackles the challenges of Section 1.3. Consequently, I expect an increase in accuracy with which documents of equivalent person entities can be identified. This can be expressed with a changed distribution in similarity values of compared documents in contrast to the situation where only recorded features are compared. Here the similarity values of non- and equivalent entities then respectively decrease and increase. Ideally this forms a clear-cut bimodal distribution with peaks for equivalent and non-equivalent entities. However, due to partially equivalent features of non-equivalent entities, I still expect a smoother distribution of similarity values.

As a case study, this research focuses identifying equivalent entities among two collections of CBG, namely Indonesian-Dutch passport requests

and personal record cards of deceased people. These both contain personal information about people and hence are valuable documents within genealogical research. However, the passport requests are currently not allowed to be freely disclosed because it is unknown whether these people already passed away (CBG, 2016e). Thus, linking these with the personal record cards will resolve this and other discussed problems from Section 1.1.

1.5 Practical and scientific contributions

To sum up, the practical contributions of this research are:

- Linking documents from different collections that belong to the same real-world people. This can be used to more accessibly and efficiently research family history, track down heirs, and disclose privacy-sensitive documents;
- Creating a data processing pipeline that can (with some adaptations) be used to accurately link documents of people (within other domains than genealogy) across collections, to documents of related people, and to more information about mentioned places;
- Specifically for CBG, linking their Indonesian-Dutch passport requests with personal record cards of equivalent, deceased people. This way they know which of the former documents they are allowed to disclose.

Moreover, the scientific contributions of this research are:

- Studying how additional information about places such as coordinates, which is not recorded within genealogical documents, influences the ability to link documents that refer to equivalent entities;
- Showing which ER-beneficial information about places can be retrieved from linked open data sources, using place names mentioned in genealogical documents;
- Combining and adapting existing similarity metrics for (genealogical) entity resolution that determine similarities of names, places, and dates;
- Discussing ethical considerations with entity resolution in general, and introducing potential biases that may occur due to name or registered sex changes;
- Facilitating meta studies that research geospatial aspects of families, such as migrations after the independence of Indonesia, by enriching data with place information and linking records from different collections. Historical trends will then also be better able to get extrapolated to the future for policy making.

The next Chapter 2 describes existing studies that perform entity resolution and linked open data within the domain of genealogy. After, Chapter 3 explains which CBG data is used and Chapter 4 the methods applied to answer the research question[2]. Next, Chapter 5 shows the resulted influence of place enrichment on equivalent entity identification, and Chapter 6 discusses these as well as future research and ethical considerations of this study. Lastly, Chapter 7 gives final answers to the research question of Section 1.4.

[2]The code created for this research can be found at the following GitHub page: https://github.com/SanderEngelberts/place-enriched-ER.

Chapter 2

Literature review

To be able to tackle the research goal of linking genealogical documents referring to equivalent people using enriched place information, Section 2.1 of this chapter first discusses entity resolution methods of contemporary research that can be combined to identify equivalent entities for the case study at CBG. After, Section 2.2 provides more information about linked open data, how such sources are structured and why it is interesting to combine this with ER.

2.1 Entity resolution for genealogical data

As introduced in Section 1.3, Entity resolution is a challenging task which tries to link documents that refer to the same real-world entity such as a person with their attributes like name, birth date, and location (Goyal, 2021). This linking can be done by calculating similarities between such available features in the documents, and combining these individual similarity values in a way that equivalent- and non-equivalent entities can get accurately distinguished (Goyal, 2021).

Specifically for genealogical data, Efremova et al. (2015) researched the linkage of people in notary deeds to their birth, marriage, and death certificates based on a person's name, location, and date. Before performing ER, to not compare each person with all other documents, they applied blocking techniques on the person's name to determine candidate pairs. Moreover, they proposed to improve on their methods by using additional context information such as a partner- or family member names to increase the linking accuracy (Efremova et al., 2015).

In contrast, Mourits et al. (2020) used a combination of the names of the key person and their partner to get more unique person identifiers for linking birth, marriage, and death certificates of the key person. They also used this and a time period threshold as blocking technique to determine candidate pairs (Mourits et al., 2020). Their research is part of the LINKS project,

which has as goal to interlink Dutch civil certificates from different archives in order to reconstruct all nineteenth and early twentieth century families in the Netherlands (Mandemakers, n.d. Mourits et al., 2020). However, unlike Efremova et al. (2015), locations were not taken into account and hence Mandemakers (n.d.) proposed to utilize this in future research to ideally allow for regional variations in the linking procedure.

Alternatively, Rahmani et al. (2014) and Rahmani et al. (2016) utilized available contextual information of all recorded relatives in a more flexible manner than Mourits et al. (2020) to determine if birth, death, and marriage certificates refer to the same person. On the one hand, Rahmani et al. (2014) checked if recorded relatives have the same blocking key and how often that one occurs within the data. On the other hand, Rahmani et al. (2016) extended this with a random walk through neighboring documents that are linked via relatives to also allow for a path via second-level (family) relationships like grandparents. These approaches increased the accuracy with which equivalent people could be identified (Rahmani et al., 2014; Rahmani et al., 2016).

Additionally, Rahmani et al. (2014) designed a blocking key string that creates distinctive blocks for Dutch names (combined with registered sex), which they and Rahmani et al. (2016) used to perform blocking. However, note that both studies determined the content similarity only based on name and registered sex similarity, which value can thus potentially be improved when more attributes such as location and date are considered like Efremova et al. (2015) did (Rahmani et al., 2014; Rahmani et al., 2016).

For identifying equivalent entities based on these location features, best a combination of place name and coordinates can be used instead of one of these individually (Sehgal et al., 2006), like Efremova et al. (2015) did with place names. This shows that enriching recorded place names with additional information about its entity can be beneficial for identifying equivalent person entities. Such added information is namely useful when multiple places exist with the same name (e.g. Laren in North-Holland and Gelderland (Alletop10lijstjes, n.d.)), when places fused, or when places have multiple (historic) name variants (e.g. Den Haag and ‘s-Gravenhage) (Sehgal et al., 2006).

2.2 Linked open data for genealogy

Additional information about places can be retrieved from linked open data sources. LOD tries to connect data sets by using the same structured formats that refer to unique entities, instead of storing these separately in relational databases with their own structures (Berners-Lee, 2009; Blaney, 2021). This way, if the same person or place is mentioned in multiple documents, then both will be referenced with the same (HTTP) Uniform Resource Identi-

fier (URI) (that is publicly available on the internet) rather than by their non-unique name string (Berners-Lee, 2009; Blaney, 2021). Then, for example, different locations with the same name will be recognized as separate entities, and a location that changed names or geometry over history has all these variants recorded within the same entity (Cuijuan et al., 2018). Retrieving such URIs that correspond to recorded place names is thus useful for identifying equivalent place entities within ER methods like the ones discussed in Section 2.1, and features of place entities can aid with gradual similarity calculations of non-equivalent places.

In more detail, a (combination of) ontologies like schema.org is used to represent relationships between an entity and information about it in a structured way (Berners-Lee, 2009; Blaney, 2021). Within genealogy this way unique people can be linked to the entities related to them such as ancestors, locations, dates, jobs, and civil certificates (Cuijuan et al., 2018). These relationships are represented (e.g. using `Turtle`) as a triple with the URI of the key person (the subject, e.g. S. Engelberts), an URI to the respective predicate from an ontology (e.g. https://schema.org/birthPlace), and the URI of the entity that is related to this person (the object, e.g. Blaricum). In such a way a full graph gets formed that can be queried with `SPARQL` on these relationships (e.g. return everyone who is born in Blaricum), but also on information that is implicitly linked due to such a structured format (e.g. a query to return everyone who is born in the Netherlands will also give S. Engelberts because Blaricum is linked with being a part of this country) (Berners-Lee, 2009; Blaney, 2021). Hence, such queries can be used to retrieve specific information about place entities to enrich the place information recorded on genealogical documents.

Chapter 3

Data

CBG made data available for the purpose of this research. As mentioned in Chapter 1, they maintain many different genealogical collections of which this research tried to interlink personal record cards from the National Register of deceased people with passport requests from the Old-Passport archive. Similar approaches as discussed in Chapter 4 can be used in future research to also link records to other collections or between people and the relatives mentioned on their documents (i.e. parent(s), child(ren), and/or partner(s)).

3.1 Personal record cards

Municipalities maintained the personal record cards of their residents between 1939 and 1994, after which the documents of living and newborn people got digitally recorded in the Personal Records Database (BRP) (CBG, 2016d). Each of the personal record cards CBG archives thus refers to a unique person entity who passed away between 1939 and 1994.

In Table 3.1 an overview is given of the attributes that are utilized in this study from these personal record cards as well as from the passport requests that are discussed in Section 3.2. Additional information that is recorded on the personal record cards is personal information about family members, marriage(s), nationality, occupation(s), living place(s) and religion. Such information can in the future be used to potentially increase the confidence further that two documents belong to an equivalent person entity.

It should be noted that often information of (at home living) children was only recorded on the personal record card of one parent, abbreviations were used, changes in for example occupations were not always recorded, not all information such as previous marriages of before 1939 were copied onto the personal record cards, and during Word War II many cards of Jewish people were destroyed (Gemeente Amsterdam Stadsarchief, n.d. Uit de oude Koektrommel, 2022).

Especially, it should be noted that at the moment of writing only meta

Table 3.1: Utilized features in personal record cards and passport requests with fictional (translated) examples of these documents. Note that not always each feature is recorded about a specific person nor digitized, and when the document never contains a certain feature then it states "Not Applicable".

Features	**Personal record card**	**Passport request**
`FirstNames`	Maria Johanna	Maria Johanna
`LastNameAffix`	Not Applicable	de
`LastName`	de Vries	Vries
`BirthDate`	1923-04-21	19230421
`BirthPlace`	Jakarta	Jakarta
`BirthCountry`	Not Applicable	Indonesia

data was available for the key person's first name, last name, birth place, and birth location. This information was retrieved from scanned cards using optical character recognition, and cleaned using multiple pre- and postprocessing procedures. After, the results were checked by volunteers via Vele-Handen.nl and where necessary corrected. This information is thus highly accurate and will in the future be appended with the remaining information on the personal record cards.

3.2 Passport requests

After the independence of Indonesia, Dutch passport requests were filed in the period 1950-1959 (CBG, 2016e). Indonesian-Dutch people who had a Dutch heritage could get granted a Dutch passport to be able to move (back) to the Netherlands. The passport requests contain valuable information for descendants because the Dutch East Indies did not keep a citizen register (CBG, 2016e). Hence, these requests can provide the information as shown in Table 3.1 about the respective ancestor and their relatives, as well as a picture of them, occupation(s), and the reason of application.

However, because this information is privacy-sensitive, these documents cannot be freely disclosed unless it is known if the respective people already passed away (CBG, 2016e). This is known after the case study of this research links these passport requests with the personal record cards.

Moreover, for this data set it should be noted that the meta data may contain many spelling errors and name variants. This is due to the labour of digitization having been done by people who do not understand Dutch and Indonesian. In the future this may be improved by domain experts using manual verification, (place) name standardization against (place) name databases, and/or optical character recognition.

Chapter 4

Methods

Multiple sequential steps are required to be able to determine which passport requests can be linked with personal record cards that refer to equivalent people, which is visualised and summarized in the flowchart of Figure 4.1. An elaboration on each of the steps from this figure can hereafter be read in their respective subsections.

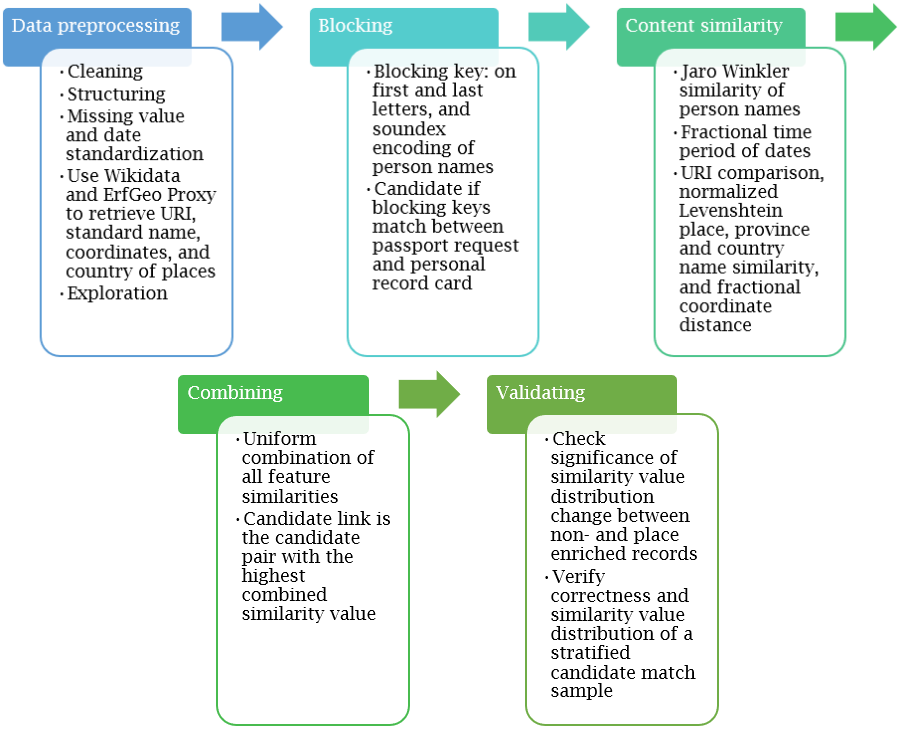


Figure 4.1: Flowchart visualising and summarizing the procedure of linking passport requests with personal record cards of equivalent people.

4.1 Data preprocessing and place enrichment

Before being able to perform entity resolution approaches, the data was collected, cleaned, and structured to be in the same format across the data sets. Missing values in attributes were standardised to be `NaN` values, and dates were changed to be in the same *YYYY-MM-DD* format. Further, partial missing dates, indicated with zeroes for one or multiple of the date elements, were also written in that format and not converted to a `NaN` value because it still contains valuable information.

Furthermore, country and Dutch province indicators were removed from the place name values using regular expressions and used to create new features. Additional text between brackets, often referring to other sections on the record, was removed as well to clean the place name values, and place name abbreviations were written out fully[1].

Next, equivalent birth places were identified at LOD sources to enrich the records with place information. For this, the python library `wptools` for querying WikiData and the ErfGeo Proxy tool for querying GeoNames on Dutch historic places were used. Whenever possible, this retrieved place entity URIs, longitude-latitude coordinates in WGS84 coordinate reference system, standardised place names, administrative region names, and country names with URI for the place names mentioned in the documents. After, the data was further explored to check how many records could get enriched with this place information, how complete the data sets are, and to check frequencies of birth places and countries. These data exploration results can be found in Sections 5.1 and 5.3.

In more detail, the ErfGeo website, which is created by among which the Cultural Heritage Agency of the Netherlands, contains toponyms (e.g. Traiecto for Maastricht), carnaval names, and dialect names that represent historic names of Dutch places (den Engelse, 2015c; RCE, n.d.). Additionally, geometries of disappeared places and of places in different time periods can be found (den Engelse, 2015c; RCE, n.d.).

Because Indonesian and other non-Dutch places can not (yet) be retrieved using ErfGeo, Wikidata was queried as well. This is a free and collaborative linked open database (Wikidata, 2022). Still, this database does not include all (Dutch) alternative names of places and (spelling) errors in the meta data of records may also result in no query results.

After a link to a place entity was created, data from the place entity was aggregated to the genealogical document themselves to get richer information (Koho et al., 2020). With data aggregation, it can occur that data from multiple sources contradicts each other, in which case a decision needs to be made about which information to display such as minimum, average,

[1]For place-, province-, and country name abbreviations and their corresponding full names see the standardized list used for personal record cards: https://cbg.nl/documents/12/Nationaal_Register_Overledenen_persoonskaarten_en_persoonslijsten.pdf

latest, or all item(s) (Knap et al., 2012). Here, the birth place name was replaced by the retrieved standardised name. Also, when a birth country name was recorded then this was replaced by the retrieved country name that corresponds to the place entity.

4.2 Blocking

Next, the entity resolution process took place. As first step, blocking was applied to reduce the number of candidate pairs[2] using Equation 4.1 that was adapted from Rahmani et al. (2014). However, registered sex (previously incorrectly referred to as 'gender') was removed from their blocking key because this was not available for the personal record cards, and can result in bias against transgender people who adapted their registered sex in legal documents.

Rahmani et al. (2014) applied blocking on Dutch names and discovered this was the best blocking key for doing so. Because the passports were requested by people with Dutch nationality or heritage, many of them have Dutch names and hence this same blocking method could be utilized in this research.

As can be seen in Equation 4.1, this blocking key string for reference r_i is the concatenation (+) of the first 3 ([: 3]) and last 2 ([−2 :]) letters of a person's first- and last name (because errors are mostly made in the middle letters), and the soundex encodings (see Appendix A) of their first- and last name (for resolving spelling errors) (Rahmani et al., 2014; Rahmani et al., 2016). Newer extensions of soundex were tried as well by Rahmani et al. (2014), but they found this version to work best for creating a blocking key on Dutch names.

$$\begin{aligned} Blocking\_key(r_i) = {} & FirstName(r_i)[:3] + FirstName(r_i)[-2:] + \\ & LastName(r_i)[:3] + LastName(r_i)[-2:] + \\ & soundex(FirstName(r_i)) + soundex(LastName(r_i)) \end{aligned} \tag{4.1}$$

Similarly, Efremova et al. (2015) performed blocking on only the phonetic encoding of Dutch names (*Double Metaphone-* and *Soundex encording*) to get candidate pairs. However, unlike Equation 4.1, this method has not been specifically proven to work well for genealogical entity resolution but instead relies on the assumption that these encodings account for different spelling variations (Efremova et al., 2015).

In contrast, the LINKS project used a different kind of blocking in which three out of four of the key person and their partner's first- and last name

[2]In this research, a candidate pair is a personal record card and passport request where the blocking key string of the key person corresponds.

needed to exactly match (Mandemakers, n.d. Mourits et al., 2020). Additionally, they only considered documents a candidate when their recorded dates were less than a lifetime apart (Mandemakers, n.d. Mourits et al., 2020). However, such a blocking key cannot be applied within this study due to partner names not yet being digitized for personal record cards. Also, this method has the challenges that people do not always have a partner or had multiple in their lifetime, and that no spelling errors or variations are allowed, unlike with Equation 4.1.

4.3 Content similarity calculation

Subsequently, the content similarity was determined between each candidate pair of which both records could be enriched with place information. In order to answer the research question of Section 1.4, this was done twice for each candidate: once without and once with the place enriched data. An example calculation of this content similarity can be seen in Appendix B.

The content similarity between two documents became the combination of three similarity values, based on the features: full person name, birth place, and birth date. Efremova et al. (2015) used these as well, instead of only focusing on a person name like Rahmani et al. (2014) and Rahmani et al. (2016), or on a person name and their partner's name like Mourits et al. (2020).

4.3.1 Person name similarity

Firstly, the Jaro Winkler similarity (see Equation A.1) between first- and last names was computed and uniformly averaged, as in Efremova et al. (2015), Rahmani et al. (2014), and Rahmani et al. (2016). This gives a gradual similarity value that allows for name variations and spelling errors, with a focus on common prefixes (Tay, 2019; Winkler, 1990). In contrast, Mourits et al. (2020) used the Levenshtein distance (see Appendix A), which also allows for a gradual similarity value.

4.3.2 Birth date similarity

Secondly, dates were compared using the fractional time period of dates (see Equation 4.2), scaled to a 10 years time interval range. In contrast, Efremova et al. (2015) created a boolean statement that checked if recorded dates were closer than 100 years, assuming this is the maximum life span of a person. However, during this study only birth dates were compared so a smaller time interval prevented high similarity scores for distant dates.

$$FractionTime(d_1, d_2, tp) = \begin{cases} 0 & \text{if } |d_1 - d_2| \geq tp \\ 1 - \frac{|d_1 - d_2|}{tp} & \text{otherwise} \end{cases} \tag{4.2}$$

where d_1 and d_2 are the compared birth dates, and tp the time period threshold. When dates were partially missing, then this equation was used on each non-missing date element pair separately, for example on years and months, and uniformly averaged. This equation is an adaptation of Equation A.3 by Geel et al. (2012), which contains the same variables, but now with higher values representing more similar dates like similarity metrics for other features.

4.3.3 Birth place similarity

Thirdly, equivalent birth places were identified based on place URI matching or four place features: place, province, and country names, and coordinates. When the Wikidata or GeoNames URIs of places were already identical, then directly the maximum place similarity value could be returned without calculating the four place feature distances (so setting these as missing values). This exemplifies the use of LOD entities when applying ER, next to being able to retrieve coordinates and other information for calculating a gradual place similarity score when place entities are not equivalent.

Specifically, Sehgal et al. (2006) found that Levenshtein distance (see Appendix A) performs best for English location name comparison of Afghan places, and hence was used here as well. This was applied for place, province, and country names, in case these were recorded or retrieved from LOD sources. In contrast, Efremova et al. (2015) returned a boolean value stating if location names were equivalent, which does not allow for name variants, spelling errors, or any of the other challenges described in Section 1.3.

Afterwards, the Levenshtein distance between place strings s_1 and s_2 was normalized to the range 0-1 using Equation 4.3. This only works well when all edit operations have cost 1, as used in this study, but alternatives exist when a dynamic program is used for computing the Levenshtein distance: for example read Fisman et al. (2022). In contrast, Sehgal et al. (2006) directly used the original Levenshtein distance value because they could learn weights between the different place similarities in a supervised way.

$$LevenshteinSim(s_1, s_2) = 1 - \frac{LevenshteinDist(s_1, s_2)}{max(length(s_1), length(s_2))} \tag{4.3}$$

Furthermore, to compare coordinates, Sehgal et al. (2006) took the inverse of the (longitude-latitude) coordinate distance $dist(l_i, l_j)$ between locations l_i and l_j (see Equation A.4). In contrast, because no supervised

learning could be used to learn appropriate feature weights, this research made an adaptation to the fractional time period Equation 4.2, as can be seen in Equation 4.4. This formula ensures that closer places get a higher similarity value than places with a larger distance, while keeping the scores within the range 0 to 1.

$$CoordSym(l_i, l_j) = \begin{cases} 0 & \text{if } dist(l_i, l_j) \geq dt \\ 1 - \frac{dist(l_i, l_j)}{dt} & \text{otherwise} \end{cases} \tag{4.4}$$

where $dist(l_i, l_j)$ is calculated using the Haversine distance by Balsebre et al. (2022) and Euclidean distance by Deng et al. (2019), and distance threshold dt is set to 200 kilometers. Which distance metric to use depends on the origin of the data, where Haversine distance is used for computing the arc distance between longitude-latitude pairs on a sphere (see Equation A.5), while Euclidean distance computes the distance on a flat plane (see Equation A.6) (Maria et al., 2020).

To avoid (large) distortions in calculated distances, the used coordinate reference system was chosen carefully (QGIS, 2020): 'Amersfoort RD/New' (EPSG:28992) for two places in the Netherlands (Kadaster, n.d.), 'Datum Geodesi Nasional 1995' (EPSG:4897) for two places in Indonesia and 'World Geodetic System 1984' (EPSG:4326) for places in different or unknown countries. The first two of these are flat planes with coordinates in meters (or similar unit), for which the Euclidean distance metric was thus most applicable. Further, the Haversine distance metric was applied for the third, spherical world CRS with longitude-latitude coordinates in degrees.

Lastly, the place, province, and country name similarity values and the coordinate similarity value were uniformly averaged (disregarding similarities between missing feature values) to determine a mean place similarity value in the range 0 to 1, with higher values representing more similar locations. As improvement, Sehgal et al. (2006) proposed to use a supervised learning approach that takes into account the skew in the training data where there are more negative than positive examples. However, during this research no labeled training data was available for this purpose.

4.4 Combining feature similarity scores and validation

As last step all components of the content similarity were uniformly averaged, giving a similarity value between 0 and 1, with higher values representing more similar person entities. Here is taken into account that missing similarity values, for example due to fully missing birth dates or person names, did not count towards the average. The example calculation in Appendix B also shows this combined similarity value calculation as well as the

decision which candidate personal record card is most likely an equivalent person entity.

Because only one of the unique personal record cards can maximally belong to each passport request, the candidate pair with highest combined similarity value was considered to be a candidate link. When the similarity value between these records reaches a (manually determined) threshold, this candidate link can be considered an actual link between records of an equivalent person entity.

To validate the results and be able to indicate a suited threshold, domain experts at CBG verified the correctness of hundred candidate links. For this, they were unaware about the calculated similarity values to reduce bias. These hundred candidate links were selected by applying stratified random sampling on the average similarity value.

Additionally, domain experts at CBG automatically linked passport requests to digital personal record cards (from Personal Records Database (BRP)) of people who passed away after 1994. They did thus using exact feature matching, where perfect corresponding matches were recorded in this test set. Hence, if this research and their linking both identified an equivalent person entity for a specific passport request, then the one of this research can be considered a false positive link.

Alternatively, like Efremova et al. (2015), weights and thresholds for each similarity value of each feature could have been learned with a supervised classification model. However, in this research no manually labeled training and test data was available which states if documents (do not) belong to the same entity. Also, on such a large collection this would be a very difficult and time consuming task to create for domain experts (Efremova et al., 2015).

Chapter 5

Results

This chapter presents the results of the enrichment of records with place information from LOD sources (Section 5.1), blocking (Section 5.2), input data exploration (Section 5.3), similarity calculations for entity resolution (Section 5.4) and the validation of its results (Section 5.5).

5.1 Enrichment of data with place information

After the data got cleaned and reformatted as described in Section 4.1, the records were enriched with place information using Wikidata and ErfGeo Proxy. Wikidata could uniquely identify a place entity URI for 9481 out of 35634 (26.6%) of the unique combinations of recorded place and province names, which ErfGeo Proxy increased to 10189 (28.6%). With respect to the separate data sets, URIs for 42.0% of the unique combinations of place and province names in personal record cards could be retrieved, whereas for the passport requests data this was only 17.6%.

As a result, 1182472 out of 1256801 (94%) personal record cards could be enriched with place information from Wikidata or ErfGeo Proxy, and 80931 out of 141382 (57%) passport requests. However, the administrative regions that were retrieved did often not refer to a province entity but also to for example municipalities or other areas, so no enrichment of province indicators could take place without manual assessment.

As a last note, not all retrieved place entities contained information about their coordinates, standard and alternative names, country name and URI, or their administrative region name and URI. Hence, during the entity resolution procedure sometimes the similarity calculation on one of these features could not be performed or was adapted in a manner as explained in Section 4.3.

5.2 Blocking

As a next step, the records that could be enriched with place information received a blocking key based on recorded first and last names, as explained in Section 4.2. This respectively resulted in 779460 and 78737 unique blocking keys for personal record cards and passport requests. Among these 12600 keys intersected, meaning that at least one personal record card was a candidate for a passport request with such a key.

Consequently, 13730 (17.0%) passport requests and 73538 (6.2%) personal record cards had at least one candidate record. As can be seen in Figure 5.1, a majority of the passport requests even had exactly one candidate personal record card to be compared with. However, in a decreasing fashion also more candidates occur per passport request.

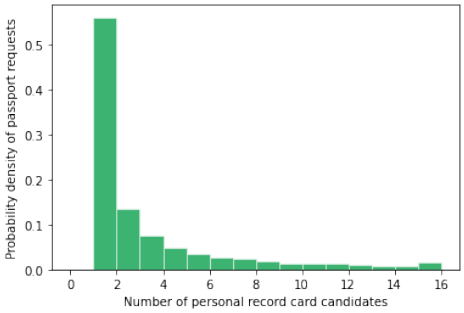


Figure 5.1: Number of candidate personal record cards per passport request after blocking, with statistics: minimum - 1, median - 1, average - 9.09, maximum - 508, total - 124859. Note that the passport requests with 0 candidates were left out of this figure, and that a majority of the remaining records have exactly one candidate.

5.3 Exploration of entity resolution input data

Only the records that could get enriched with place information and that have at least one candidate after blocking were considered as input for the entity resolution procedure. This made it possible to conclude if place enrichment increases the accuracy with which equivalent person entities can be identified, while increasing the algorithmic efficiency by only comparing candidate records that are more likely to be equivalent.

Among the input data there was a negligible number of missing values of person names, birth dates and place entity URIs. Additionally, there were no missing birth place names because these were used to retrieve their unique place URI. However, almost none of the passport requests contained a Dutch province name indicator in its originally recorded birth place name,

so its corresponding string similarity could rarely be calculated.

Furthermore, around half of the previously missing birth country names could be filled in due to place enrichment, resulting in recorded country names for 84.3% of the personal record cards and 58.3% of the passport requests. As to be expected with Dutch records, most of the recorded countries equaled the Netherlands, namely 95.7% of the personal record cards and 83.4% of the passport requests. Further, Dutch-Indonesian people who requested a passport needed to be of Dutch heritage but did not have to be born there, so also 7.2% of these recorded countries equaled Indonesia, whereas this was only 2.3% of the personal record cards. This frequency difference can also be seen in the most common place names, where there are more Indonesian place names among common birth places of passport request than of personal record cards.

Lastly, for personal record cards 18673 (25.4%) coordinate pairs could be retrieved, and for passport requests 5658 (41.2%). Thus, for the candidates that did not have an equivalent place URI, some could get a gradual place similarity that was partially based on geographical closeness.

5.4 Content similarity calculation

After executing the entity resolution procedure as described in Section 4.3, 14172 (103.22%) candidatelinks were found for non-place enriched data and 14444 (105.20%) for place enriched data. Of these, 12236 (89.12%) candidate links overlapped. These are more than the 13730 inputted passport requests because some candidates had the same maximum average similarity and hence were all returned as candidate link. The similarity value statistics of these candidate links are shown in Table 5.1 and Figure 5.2.

As can be seen in Table 5.1, the average similarity and the average place similarity increase with place enrichment. However, this change does not directly imply an improvement of accuracy so that will be verified in Section 5.5 hereafter.

On the one hand, some individual place similarity scores got lower averages, which can possibly be explained by their increased number of missing values. This latter value increased due to place URIs that already exactly matched between records so these metrics only being calculated for places which are non-equivalent. On the other hand, higher individual place similarity scores were retrieved when less missing values occurred due to place enrichment.

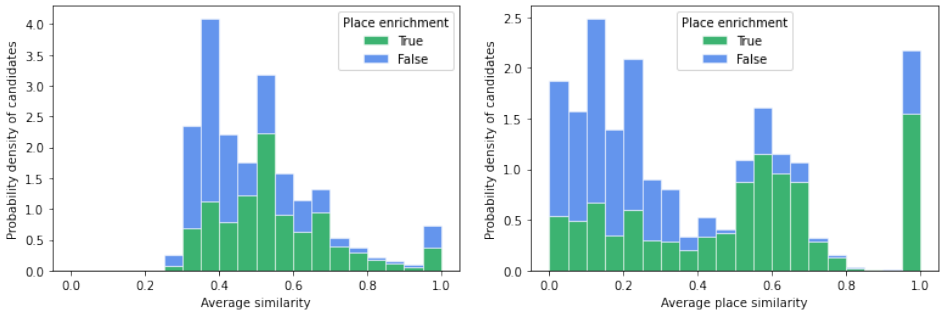
Moreover, the person name similarity values are very high due to blocking. However, this does not directly mean that a candidate is also an equivalent entity, which for example the low average birth date similarity value shows.

Lastly, Figure 5.2 visually shows that there is a difference in the average

Table 5.1: Statistics of similarity scores between passport requests and candidate personal record cards. If multiple numbers are mentioned, separated by a semicolon, then the first and second numbers respectively represent the similarity score from non- and place enriched data. If only one number is mentioned, then the value is the same in both situations. Note that place enrichment increases the average (place) similarity, but also has different influences on the individual place similarity scores.

Similarity	**Minimum**	**Mean (std)**	**Maximum**	**Number of missing values**
Mean place, birth date, and person name	0.22	0.46 (±0.16) ; 0.55 (±0.16)	1.00	0
Mean place	0.00	0.26 (±0.26) ; 0.51 (±0.30)	1.00	0
Place name	0.00	0.22 (±0.24) ; 0.16 (±0.13)	1.00	0 ; 19254
Country name	0.00	0.79 (±0.32) ; 0.91 (±0.24)	1.00	107033 ; 61448
Province name	0.00	0.48 (±0.36) ; 0.44 (±0.35)	1.00	123368 ; 123581
Coordinates	NaN ; 0.00	NaN ; 0.60 (±0.28)	NaN ; 1.00	124859 ; 67002
Birth date	0.00	0.17 (±0.31)	1.00	24
Mean person name	0.65	0.96 (±0.05)	1.00	0
First name	0.56	0.97 (±0.06)	1.00	61
Last name	0.55	0.96 (±0.07)	1.00	0

(place) similarity score distributions when records were (not) enriched with place information. This is not a clear-cut bimodal distribution with peaks for non- and place enriched candidates as would be the ideal situation. Instead, most candidates seem to have scores lower than 0.8 and there is a peak around 0.95 to 1.0 again.



(a) Average similarity (b) Average place similarity

Figure 5.2: Probability density distributions of similarity scores between passport requests and their candidate personal record cards. These figures display differences in the score distributions between records that were (not) enriched with place information.

5.5 Validation

Firstly, the average (place) similarity distributions are indeed significantly different between non- and place enriched candidates, as a Wilcoxon signed-rank test indicated[1]. This test was used instead of a t-test because a Z-test indicated the existence of outliers and a Shapiro test indicated that the distributions are non-Gaussian (as can also be visually concluded from Figure 5.2), and hence its assumptions were violated.

Secondly, domain experts at CBG validated a stratified random sample of hundred candidate links and matched the passport requests against the BRP records of deceased people. This resulted in the true- and false positive distributions for average (place) similarity values that can be seen in Figure 5.3[2], and its statistics as shown in Table 5.2.

In specific, 3447 (25.11%) passport requests that have minimum one candidate personal record card could be exactly matched with a BRP record of a person who passed away after 1994. Thus, the candidate links of these passport requests were considered false positives.

Furthermore, for the manually validated sample only the candidate links that exceeded the following thresholds were considered: average name similarity ≥ 0.90, birth date similarity ≥ 0.5, and average place similarity ≥ 0.5. These values were selected based on Table 5.1 and the resulting average similarity distribution of the considered candidate links. This has been done due to the majority of passport requests only having one candidate personal record card, so not all returning a reasonable candidate link for validation.

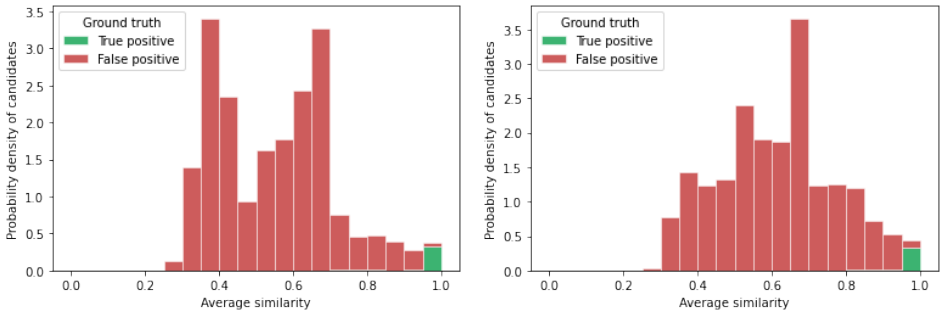
[1]The Wilcoxon signed-rank test results when pairwise comparing the difference between similarity scores of non- and place enriched data are respectively for average similarity and average place similarity: 61541503.0 (p-value 0.0) and 61545811.5 (p-value 0.0).

[2]Inspired by Figure 4 of Rahmani et al. (2014).

As a result, 33% of the sampled candidate links were false positives and 67% true positives.

Thirdly, in contrast to the hypothesis of Section 1.4, Table 5.2 shows that both the true- and false positives got higher average (place) similarity scores after enriching the data with place information. Still, true positives in general have higher scores and a lower standard deviation than false positives, as can also be inferred from Figure 5.3. A threshold for links can thus also be set very high for both non- and place enriched data: around 0.95 average similarity. However, false negatives also have some scores in that region.

Lastly, a Wilcoxon signed-rank test indicated that the average (place) similarity distributions are significantly different between non- and place enriched candidate links for false positives, but not for true positives. In Table 5.2 can indeed be seen that true positives already had high values so place enrichment did not increase these values a lot, in contrast to the values of false positives.



(a) Non-place enriched

(b) Place enriched

Figure 5.3: Probability density distributions of average similarity scores between candidate links of the validated data. These figures display differences in the score distributions between false positive and true positive candidate links for non- and place enriched data. Here can also be seen that true positives have in general higher scores and a lower standard deviation than false positives. Further, the distribution of false positive values looks different between the figures.

Table 5.2: Statistics of similarity scores between true- and false positive candidate links of the validated data. If multiple numbers are mentioned, separated by a semicolon, then the first and second numbers respectively represent the similarity score from non- and place enriched data. If only one number is mentioned, then the value is the same in both situations. No maximum values are added in this table because these are the same as in Table 5.1. Note that true positives have higher scores than false positives, and that place enrichment in general increased the average- and minimum (place) similarity values for both.

Similarity	**True positives mean (std)**	**False positives mean (std)**	**True positives minimum**	**False positives minimum**
Mean place, birth date, and person name	0.98 (±0.05) ; 0.99 (±0.03)	0.54 (±0.15) ; 0.61 (±0.15)	0.26 ; 0.85	0.26
Mean place	0.97 (±0.14) ; 0.99 (±0.07)	0.38 (±0.33) ; 0.58 (±0.32)	0.18 ; 0.55	0.00
Place name	0.97 (±0.13) ; 0.78 (±0.31)	0.34 (±0.33) ; 0.21 (±0.15)	0.18 ; 0.29	0.00
Country name	0.87 (±0.26) ; 0.70 (±0.52)	0.80 (±0.32) ; 0.90 (±0.26)	0.48 ; 0.10	0.09
Province name	NaN	0.64 (±0.38) ; 0.57 (±0.38)	NaN	0.00
Coordinates	NaN ; 0.93 (±0.10)	NaN ; 0.62 (±0.29)	NaN ; 0.81	NaN ; 0.00
Birth date	0.99 (±0.04)	0.29 (±0.35)	0.80	0.00
Mean person name	1.00 (±0.01)	0.96 (±0.05)	0.94	0.67
First name	0.99 (±0.02)	0.96 (±0.07)	0.89	0.59
Last name	1.00 (±0.01)	0.97 (±0.06)	0.93	0.58

Chapter 6

Discussion

Now the results of this study are shown in Chapter 5, these will first be discussed in Section 6.1. After, ethical considerations of ER and LOD in general, and this research in specific are examined in Chapter 6.2. Lastly, mainly based on these two chapters, future research directions are proposed in Chapter 6.3.

6.1 Results discussion

Before being able to draw conclusions and propose future research, the results of Chapter 5 need to be further investigated. This will be done in the subsections hereafter that correspond to the sections with the same numbers and titles as in Chapter 5.

6.1.1 Enrichment of data with place information

A difference could be seen between how many passport requests and how many personal record cards could get enriched with place information from LOD sources. There are multiple likely reasons for this:

- The passport requests were digitized by people who do not understand Dutch or Indonesian, so contain many spelling errors. The place URIs were retrieved for exact place name matches, so cannot retrieve place entities for faulty names.
- There are more spelling variations of place names among passport requests because of the lack of standardization (of Indonesian place names). For example, the place name variants “Soerabaia”, “Surabaia”, and “Soerabaja” all exist within the data. In contrast, personal record cards only contain standard place name variants of Dutch places. Not all of these (Dutch) name variants of Indonesian places are recorded in Wikidata, and non-Dutch places do not exist in ErfGeo Proxy.

- Almost none of the passport requests contained additional Dutch province indicators, unlike the personal record cards. These could have been used to resolve conflicts between places with the same name that are located in different provinces.

6.1.2 Blocking

Only a part of the passport requests had at least one candidate personal record card based on similar personal names. The full personal record data set was not available for this study yet, so (around 4 times) more passport requests will likely get at least one candidate in future research. Also, not everyone that filed a passport request will have passed away before 1994 so no equivalent personal record card should exist for them. Again, there may also be errors in the digitization of personal names on passport requests, but phonetic encoding and only taking first and last letters into account for the blocking key should have accounted for this in some degree. Additionally, domain experts at CBG indicated that they also found some inconsistencies in passport requests where last names of partners were used instead of maiden names like in personal record cards. This will have also resulted in wrong or no candidates for these records.

Furthermore, Figure 5.1 showed that a majority of the remaining passport requests have only one candidate personal record card, which results in that card directly becoming the candidate link. Lower average similarity scores for candidate links, as seen in Figure 5.2, can thus be explained by this. False positive candidate links in that low similarity score region, as seen in Figure 5.3, can be explained in a similar fashion.

6.1.3 Exploration of entity resolution input data

Place enrichment could not retrieve additional information of each feature, which impacted the individual place metrics that could be calculated during entity resolution. Hence, when either of the candidates did not have a feature recorded, then this feature was not considered within the average place similarity value.

However, these missing values will have posed a lesser issue for place enriched data because firstly less records retained missing features, and secondly these features were not considered if place URIs already equaled. Especially coordinates were interesting to utilize for a more gradual place similarity metric when the URIs do not match, as discussed by Sehgal et al. (2006), but only a minority of the passport requests got these with place enrichment.

6.1.4 Content similarity

Firstly, many candidate links were equal between non- and place enriched data, which makes sense with the high number of passport requests that only had one candidate personal record card. This one candidate was often not an equivalent entity, as can be seen with, for example, the low average birth date similarity value.

Secondly, the person name similarities have high values due to the blocking procedure. Also, the average country name similarities are high due to many birth places being in the Netherlands, so for different places also being equal. For the place enriched data this is likely higher due to less missing values, and due to replacing multiple recorded birth countries by the one belonging to the retrieved birth place entity.

Thirdly, remaining missing values within place features were due to place URIs already matching so these not having been calculated or either of the records not containing this information. Because the place feature scores of equal place entities were not taken into consideration, the average place and province name similarity values decreased accordingly. If these feature similarities were directly set to the maximum value instead of a missing value, then these averages are likely much higher.

Lastly, the increase in average (place) similarity after place enrichment may partly be due to the introduced average coordinate similarity being higher than the average (place) similarity score of non-place enriched data. Hence, it is important that a sample of the results was manually validated to determine if place enrichment improves equivalent entity identification.

6.1.5 Validation

Enriching the data with place information does not seem to aid in the identification of equivalent entities when uniform averages are taken, but just changes the similarity score distributions of non-equivalent entities. Equivalent entities already had high scores without place enrichment thus did not benefit as much from this extra information as false positive candidate links, although these still got lower standard deviations. Different thresholds for the gradual birth date and coordinate similarity metrics can influence how well false- and true positive candidate links can be separated, as well as different weighting of each individual similarity metric within the average similarity scores. Calibrating these to different values is thus a point of focus for future proposed research in Section 6.3.

6.2 Ethical considerations

Performing data science studies often comes with ethical considerations, which this section shows. Firstly, Binette and Steorts (2020) express that it

is important to use case studies such as this research to be able to evaluate how well the methods work in practice within the intended domain. However, because the used data is privacy sensitive, not all researchers would be able to use this same data set for research duplication or comparison, in contrast to open source benchmark data sets. This should not pose a problem due to the publication of this paper and its accompanying code. Still, it may be interesting to synthesize and publish fake data that has similar distributions as the real data, which can be used by researchers to design improved ER algorithms (Qin et al., n.d.).

Secondly, a blocking procedure, like explained in Section 4.2, is required for efficiently scaling to large data sets but can also introduce bias against certain (minority/ oppressed) groups. Depending on the data, consider here for example people who changed their first name, took the last name of their partner after marriage, or adapted their registered sex. Corresponding feature similarity metrics could then also return low values while the entities are equivalent. This is especially problematic for identifying equivalent entities among collections that are from different time periods where there is more likelihood such a change occurred in the mean time, or that do not have to use feature values from legal documents such as prayer cards. For this research documents were used that record the legal first name and maiden name so name changes did not pose a big problem, however can still have resulted in discussed bias.

For changing legal (first or last) names in the Netherlands and Indonesia, a petition at the court should be made by a lawyer with valid reasons (MoFA, 2018; RSC, n.d.). Since 1985 this is also possible for changing the registered sex in the Netherlands, under certain procedures, where a historic trace of this change is kept within the personal record cards and an adaptation is made within all other documents and within excerpts for other institutions (JenV, 2014). Hence, if all changes within personal record cards are digitized, then this bias can be alleviated by creating multiple (person name) blocking keys for adapted records when comparing against this collection. During this study however, such additional information was not available so some equivalent entities may have not been identified while these existed.

Thirdly, entity resolution and linked open data pose privacy considerations due to their nature of linking multiple data sources (Binette & Steorts, 2020; Corsar et al., 2013). (Unlawful) disclosure or malicious use of this data becomes a more serious risk because more information about individuals can get leaked (Binette & Steorts, 2020; Corsar et al., 2013). For instance, this data can be used to de-anonymize other, unrelated data sets (Binette & Steorts, 2020). There thus needs to be even more careful protection of the data to prevent (unlawful) privacy breaches (Binette & Steorts, 2020; Corsar et al., 2013). This research for that purpose and for data deduplication only saved the document numbers belonging to equivalent entities instead of recording all linked data within one source.

Lastly, Section 1.5 already highlighted some societal and scientific promises this research has, for example: the disclosure of passport requests because it is now known that the respective person passed away already, researching potential improvements of entity resolution procedures within genealogy, and enriched data that can aid in family history (meta) studies. For the first example it is of course important to not have any false positive links, so manual validation of the final results may be required. If this is not feasible, then still the similarity values can be displayed to transparently inform the users of the CBG archive what a link between a passport request and a personal record card is based on and how certain the algorithm is about it.

6.3 Future research

Firstly, in the future, additional manual cleaning and lookup may be required to retrieve more equivalent place entities or decide among multiple LOD query results (den Engelse, 2015a), and to be able to utilize the retrieved administrative regions information. Additionally, different sources may be utilized to retrieve values for missing attributes using the unique place URIs. Also, for CBG it is still interesting to run entity resolution on records that could not be enriched with place information, but this may have lower certainty and different thresholds of identifying equivalent entities.

Secondly, it is interesting to experiment with the use of a different blocking key when no or only few candidates were found, or when no equivalent entity was identified with the original blocking key. This to make sure this is not due to spelling errors or personal name changes. An example key can include different features such as (birth) dates, like Mourits et al. (2020), and (birth) places.

Thirdly, supervised learning methods, like for example discussed in Binette and Steorts (2020), can best be used to learn optimal thresholds and weights for each feature similarity value. This in contrast to iterative manual calibration and validation. Here then no normalization and distance/ time period thresholds need to be taken into account, because the weights will take this into account. Optimized weights can then possibly still result in the hypothesized increased ability to correctly split the candidate links into non- and equivalent entities, in contrast to this study.

As training and test data for this proposed procedure, manually validated candidate(s) (links) of this research can be used. This is a relatively feasible number of candidate records to verify, whereas manually labeling with the full data is too complicated as discussed by Efremova et al. (2015) and hence not used for this study either. Additionally, only candidates will be labeled this way, which are the ones that will be compared due to the blocking mechanism in the algorithm. Important to note is that class imbalances should be carefully tackled due to a small number of candidates referring to

equivalent entities.

As extension, Geographically Weighted Regression or other geospatial methods can be used to verify if spatial non-stationarity exists among feature similarity scores (Brunsdon et al., 1996). If this is not the case, then a standard global model can be applied that determines if two entities are equivalent based on learned weights and thresholds. Otherwise, using such local geospatial methods, different weights and thresholds can be determined for different regions or varying distances between place entities. Likewise, Mandemakers (n.d.) proposed to take into account regional differences within ER, as well as temporal variations.

Fourthly, taking into account the frequency of (person or place) names can ensure that entities with less common names are considered more likely equivalent than ones with more common names (Binette & Steorts, 2020). For person names this can also be done using the number of candidates for a record, similarly to how Rahmani et al. (2014) did this for relatives.

Fifthly, different features that are not yet digitized in this data may further aid identifying equivalent entities, and make it possible to interlink documents of relatives that are mentioned on these records. This can be done in a similar fashion as this research, where the values of the feature are first matched against a LOD source to resolve name variations and enrich the data with additional attributes. In specific, Rahmani et al. (2014) and Rahmani et al. (2016) have shown that context similarity increases the ER accuracy. A proposal how this information about recorded relatives can be used can be read in Appendix C.

Lastly, implementing a LOD ontology for CBG would represent each document in all collections in a standardized format, which enriches the data through its relationships with other entities. This way, it is possible to infer new information about, for example, the siblings or grandparents of someone instead of only about their parents, children or partners that are directly mentioned in their records (Rahmani et al., 2014; Rahmani et al., 2016). Hence, in the future a full family tree can be easily inferred that refers to all documents related to each ancestor, and (meta) studies about geospatial aspects of families are facilitated.

For such a LOD ontology it is interesting to explore the use of time-dependent predicates, where changes in for example marriage, person names, and registered sex are recorded with a time stamp. Consequently, entity resolution procedures can identify equivalent entities for documents based on the time these were created in, with among others no more bias against transgender people. When this LOD ontology is created, the outputs of this research can also be represented in this format by creating relationships between person entities and their related documents. An example of this procedure is given in Appendix D. When LOD entities are then not only used as input for entity resolution, but its results are also presented in such a way, then the data is optimally used.

Chapter 7

Conclusion

This research attempted to improve entity resolution algorithms within the genealogical domain by enriching documents with place information. As first step, this study showed that external information about places can be retrieved in an automated manner from linked open data sources such as Wikidata or GeoNames that refer to unique entities. However, additional manual cleaning and lookup can still increase the number of place entities that are retrieved. With this enriched data it is possible to confidently state if two place entities are exactly equal, and otherwise calculate a more intricate similarity metric between their retrieved features such as coordinates, country names, and standardized place names.

Moreover, Sehgal et al. (2006) have shown that a combination of place name (with Levenshtein distance) and coordinates (with inverse coordinate distance) can better identify equivalent places than one of these alone. Specifically for calculating coordinate similarities, it is important to limit distance distortions by converting coordinates to a suitable coordinate reference system, in which the retrieved country identifiers play a role, and use an appropriate distance metric for that projection. Further, this study has adapted these metrics to be normalized between 0 and 1 in order to uniformly weight features. As extension, future studies can validate the resulting candidate links of this research to form a training and test data set for supervised learning of optimal feature weights and thresholds.

To conclude, with uniform weighting of person name, birth date and birth place features this research was not able to show a significant benefit of enriching genealogical documents with place information for identifying equivalent person entities. However, significant similarity value distribution changes were found, specifically for non-equivalent entities. Hence, with different weighting methods proposed future studies may still be able to prove an increased accuracy. This then continues to expand ER research, improve efficiency in family history (meta) studies, and allows disclosure of more passport requests at CBG.

Chapter 8

Acknowledgements

Diving into the niche world of genealogy research with real-world data was a very interesting and fun experience for me. I am grateful that I was given the opportunity to apply my data science skills within this domain by studying novel ways of enriching entity resolution procedures using linked open data sources. I believe that the combination of entity resolution and linked open data is a valuable future direction of this field that can give many scientific and practical contributions. Although my research did not achieve the hypothesized results, I could show new methods that build on top of previous studies, discuss potential shortcomings with its future improvements, and explore the ethical considerations that need to be taken into account within this field.

For now, I want to take the time to thank some important people who helped me reach this point. First of all, thank you very much Toine Pieters and Simon Dirks for our meetings and your written feedback. This helped me to make sure my thesis is scientifically relevant for and in the scope of a Applied Data Science master thesis at Utrecht University. Also a big thank you to Pieter Woltjer for giving me the opportunity to perform this research at CBG and guide me with your domain experience in this interesting field. This showed me the practical relevance of my thesis and gave me interesting insights into the data and the workfield. Moreover, I want to thank Pieter Woltjer and Jeroen Balkenende for validating my results such that this could be done in an unbiased fashion by domain experts. Next, I want to thank Nick Peters, my fellow master student on this topic, for the talks about our theses and giving each other an insight into different aspects of this domain. And last but definitely not least, I want to thank my partner Florian Kühnert, family and friends for always being there for me to give emotional support and listen to my random monologues on this topic when I figured out something new. And not to forget, thank you reader for your interest in my thesis. I hope I could inspire you to (continue) pursue a project in this interesting area.

Bibliography

Alletop10lijstjes. (n.d.). 10 nederlandse plaatsen met dezelfde naam [https://www.alletop10lijstjes.nl/10-nederlandse-plaatsen-met-dezelfde-naam/].

Balsebre, P., Yao, D., Cong, G., & Hai, Z. (2022). Geospatial entity resolution. *Proceedings of the ACM Web Conference 2022*, 3061–3070. https://doi.org/10.1145/3485447.3512026

Berners-Lee, T. (2009). Linked data [https://www.w3.org/DesignIssues/LinkedData.html].

Binette, O., & Steorts, R. C. (2020). (almost) all of entity resolution. https://doi.org/10.48550/ARXIV.2008.04443

Blaney, J. (2021). Introduction to the principles of linked open data (A. Crymble, Ed.). https://doi.org/10.46430/phen0068

Brunsdon, C., Fotheringham, A. S., & Charlton, M. E. (1996). Geographically weighted regression: A method for exploring spatial nonstationarity. *Geographical Analysis*, *28*(4), 281–298. https://doi.org/https://doi.org/10.1111/j.1538-4632.1996.tb00936.x

CBG|Centrum voor familiegeschiedenis. (2016a). CBG verzamelingen [https://cbg.nl/bronnen/cbg-verzamelingen/].

CBG|Centrum voor familiegeschiedenis. (2016b). CBG verzamelingen zoekuitleg [https://cbgverzamelingen.nl/zoekuitleg].

CBG|Centrum voor familiegeschiedenis. (2016c). Erfgenamenonderzoek [https://cbg.nl/diensten/voor-notarissen/].

CBG|Centrum voor familiegeschiedenis. (2016d). Nationaal Register Overledenen (persoonskaarten en -lijsten) [https://cbg.nl/bronnen/cbg-verzamelingen/persoonskaarten-en-lijsten/].

CBG|Centrum voor familiegeschiedenis. (2016e). Oost-Indische bronnen [https://cbg.nl/bronnen/cbg-verzamelingen/oost-indische-bronnen/].

CBG|Centrum voor familiegeschiedenis. (2016f). Over CBG: Geschiedenis [https://cbg.nl/over-het-cbg/geschiedenis/].

CBG|Centrum voor familiegeschiedenis. (2016g). Over CBG: Organisatie [https://cbg.nl/over-het-cbg/organisatie/].

Corsar, D., Edwards, P., & Nelson, J. (2013). Personal privacy and the web of linked data. In S. Decker, J. Hendler, & S. Kirrane (Eds.),

Proceedings of workshop on society, privacy and the semantic web - policy and technology (privon2013) (pp. 1–11). CEUR-WS.

Cuijuan, X., Wei, L., & Lei, Z. (2018). Implementation of a linked data-based genealogy knowledge service platform for digital humanities. *Data and Information Management*, *2*(1), 15–26. https://doi.org/10.2478/dim-2018-0005

D'Agostino, M., & Dardanoni, V. (2009). What's so special about euclidean distance? a characterization with applications to mobility and spatial voting. *Social Choice and Welfare*, *33*, 211–233. https://doi.org/10.1007/s00355-008-0353-5

den Engelse, M. (2015a). Handleiding OpenRefine / ErfGeoProxy [https://erfgeo.nl/wat-hoe/openrefine.html].

den Engelse, M. (2015b). Standaardiseren [https://erfgeo.nl/wat-hoe/standaardiseren.html].

den Engelse, M. (2015c). Wat voor data vind je in ErfGeo? [https://erfgeo.nl/wat-hoe/watvoordata.html].

Deng, Y., Luo, A., Liu, J., & Wang, Y. (2019). Point of interest matching between different geospatial datasets. *ISPRS International Journal of Geo-Information*, *8*(10). https://doi.org/10.3390/ijgi8100435

Efremova, J., Calders, T., & Weiss, G. (2015). Multi-source entity resolution for genealogical data. *Population Reconstruction*, 129–154. https://doi.org/10.1007/978-3-319-19884-2_7

Ehrmann, M., Hamdi, A., Pontes, E. L., Romanello, M., & Doucet, A. (2021). Named entity recognition and classification on historical documents: A survey. *CoRR*, *abs/2109.11406*. https://arxiv.org/abs/2109.11406

Fisman, D., Grogin, J., Margalit, O., & Weiss, G. (2022). The normalized edit distance with uniform operation costs is a metric. https://doi.org/10.48550/ARXIV.2201.06115

GDPR.eu. (2022). Recital 27: Not applicable to data of deceased persons [https://gdpr.eu/recital-27-not-applicable-to-data-of-deceased-persons/].

Geel, M., Church, T., & Norrie, M. C. (2012). Mix-n-match: Building personal libraries from web content. In P. Zaphiris, G. Buchanan, E. Rasmussen, & F. Loizides (Eds.), *Theory and practice of digital libraries* (pp. 345–356). Springer Berlin Heidelberg.

Gemeente Amsterdam Stadsarchief. (n.d.). Persoons- en archiefkaarten 1939-1994 [https://archief.amsterdam/uitleg/indexen/28-persoons-en-archiefkaarten].

Goyal, S. (2021). An introduction to entity resolution — needs and challenges [https://towardsdatascience.com/an-introduction-to-entity-resolution-needs-and-challenges-97fba052dde5].

Haldar, R., & Mukhopadhyay, D. (2011). Levenshtein distance technique in dictionary lookup methods: An improved approach. https://doi.org/10.48550/ARXIV.1101.1232

Kadaster. (n.d.). Rijksdriehoeksmeting (RD) [https://www.kadaster.nl/zakelijk/registraties/basisregistraties/rijksdriehoeksmeting].

Kassiri, A. E., & Belouadha, F.-Z. (2017). A foaf ontology extension to meet online social networks presentation and analysis. *2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI)*, 3056–3061. https://doi.org/10.1109/ICPCSI.2017.8392287

Knap, T., Michelfeit, J., & Necaský, M. (2012). Linked open data aggregation: Conflict resolution and aggregate quality. *2012 IEEE 36th Annual Computer Software and Applications Conference Workshops*, 106–111. https://doi.org/10.1109/COMPSACW.2012.29

Koho, M., Leskinen, P., & Hyvönen, E. (2020). Integrating historical person registers as linked open data in the warsampo knowledge graph. In E. Blomqvist, P. Groth, V. de Boer, T. Pellegrini, M. Alam, T. Käfer, P. Kieseberg, S. Kirrane, A. Meroño-Peñuela, & H. J. Pandit (Eds.), *Semantic systems. in the era of knowledge graphs* (pp. 118–126). Springer International Publishing.

Leskinen, P., & Hyvönen, E. (2020). Linked open data service about historical finnish academic people in 1640–1899. *Proceedings of Digital Humanities in Nordic Countries (DHN 2020), Riga*. http://ceur-ws.org/Vol-2612/short14.pdf

Leskinen, P., Tuominen, J., Heino, E., & Hyvönen, E. (2017). An ontology and data infrastructure for publishing and using biographical linked data [International Semantic Web Conference, ISWC ; Conference date: 21-10-2017 Through 25-10-2017]. In A. Adamou, E. Daga, & L. Isaksen (Eds.), *Proceedings of the second workshop on humanities in the semantic web (whise ii)* (pp. 15–26). CEUR. http://ceur-ws.org/Vol-2014/paper-02.pdf

Mandemakers, K. (n.d.). LINKS: LINKing system for historical family reconstruction [https://iisg.amsterdam/files/2018-10/hsn_projects_links-project.pdf].

Maria, E., Budiman, E., Haviluddin, & Taruk, M. (2020). Measure distance locating nearest public facilities using haversine and euclidean methods. *Journal of Physics: Conference Series*, *1450*(1), 012080. https://doi.org/10.1088/1742-6596/1450/1/012080

Meroño-Peñuela, A., Ashkpour, A., Guéret, C., & Schlobach, S. (2016). Cedar: The dutch historical censuses as linked open data. *Semantic Web*, *8*, 1–14. https://doi.org/10.3233/SW-160233

Ministerie van Justitie en Veiligheid. (1998). Burgerlijk wetboek boek 1, artikel 17a [http://wetten.overheid.nl/jci1.3:c:BWBR0002656&boek=1&titeldeel=4&afdeling=2&artikel=17a].

Ministerie van Justitie en Veiligheid. (2014). Informatieblad wet wijziging vermelding van het geslacht in de geboorteakte (transgenders) [https://www.rijksoverheid.nl/documenten/brochures/2014/06/20/informatieblad-wet-wijziging-vermelding-van-het-geslacht-in-de-geboorteakte-transgenders].

Ministry of Foreign Affairs Republic of Indonesia. (2018). Changes in passport details [https://kemlu.go.id/doha/en/pages/perubahan_data_paspor/946/about-service].

Mourits, R., van Dijk, I. K., & Mandemakers, K. (2020). From matched certificates to related persons. *Historical Life Course Studies*, *9*, 49–68. https://hdl.handle.net/10622/23526343-2020-0006

Nurmikko-Fuller, T., Dix, A., Weigl, D. M., & Page, K. R. (2016). In collaboration with in concert: Reflecting a digital library as linked data for performance ephemera. *Proceedings of the 3rd International Workshop on Digital Libraries for Musicology*, 17–24. https://doi.org/10.1145/2970044.2970049

Pellissier Tanon, T., Weikum, G., & Suchanek, F. (2020). Yago 4: A reasonable knowledge base. In A. Harth, S. Kirrane, A.-C. Ngonga Ngomo, H. Paulheim, A. Rula, A. L. Gentile, P. Haase, & M. Cochez (Eds.), *The semantic web* (pp. 583–596). Springer International Publishing.

Pine, L. G. (2021). Encyclopedia Britannica: Genealogy [https://www.britannica.com/topic/genealogy].

QGIS. (2020). Coordinate reference systems [https://docs.qgis.org/3.4/en/docs/gentle_gis_introduction/coordinate_reference_systems.html].

Qin, X., Chai, C., Tang, N., Li, J., Luo, Y., Li, G., & Zhu, Y. (n.d.). Synthesizing privacy preserving entity resolution datasets.

Rahmani, H., Sahraei, B. R., Weiss, G., & Tuyls, K. (2014). Contextual entity resolution approach for genealogical data. *LWA*. https://www.semanticscholar.org/paper/Contextual-Entity-Resolution-Approach-for-Data-Rahmani-Sahr%20aei/8b9dd36da10f9458bbf72d42277b1a8810139304

Rahmani, H., Ranjbarsahraei, B., Weiss, G., & Tuyls, K. (2016). Entity resolution in disjoint graphs: An application on genealogical data. *Intelligent Data Analysis*, *20*(2), 455–475. https://doi.org/10.3233/IDA-160814

Rechtspraak Servicecentrum. (n.d.). Wijzigen voornaam [https://www.rechtspraak.nl/Onderwerpen/Wijzigen-voornaam].

Rijksdienst voor het Cultureel Erfgoed. (n.d.). Erfgeo [https://www.cultureelerfgoed.nl/onderwerpen/bronnen-en-kaarten/overzicht/erfgeo].

Schema.org. (2022). Relatedlink: A schema.org property [https://schema.org/relatedLink].

Schermer, B. W., Hagenauw, D., & Falot, N. (2018). Handleiding algemene verordening gegevensbescherming [https://autoriteitpersoonsgegeve

ns.nl/sites/default/files/atoms/files/handleidingalgemeneverordeninggegevensbescherming.pdf]. *Ministerie van Justitie en Veiligheid*.

Sehgal, V., Getoor, L., & Viechnicki, P. D. (2006). Entity resolution in geospatial data integration. *Proceedings of the 14th Annual ACM International Symposium on Advances in Geographic Information Systems*, 83–90. https://doi.org/10.1145/1183471.1183486

Tay, K. (2019). What is jaro/jaro-winkler similarity? [https://statisticalodds andends.wordpress.com/2019/09/11/what-is-jaro-jaro-winkler-similarity/].

Uit de oude Koektrommel. (2022). Uitleg persoonskaart [https://www.uitdeoudekoektrommel.com/uitleg-persoonskaart/].

van Koutrik, V., & Welings, Y. (2019). Beperkt waar het moet: Handreiking voor het stellen van beperkingen aan de openbaarheid van overheidsinformatie bij overbrenging naar een archiefbewaarplaats onder de Archiefwet 1995 [https://vng.nl/sites/default/files/2019-11/beperkt-waar-het-moet_20190726.pdf]. *Vereniging van Nederlandse Gemeenten*.

Vykhovanets, V. S., Du, J., & Sakulin, S. A. (2020). An overview of phonetic encoding algorithms. *Automation and Remote Control*, *81*, 1896–1910. http://valery.vykhovanets.ru/Texts/2020/Vykhovanets2020_1.pdf

Wikidata. (2022). Wikidata:introduction [https://www.wikidata.org/wiki/Wikidata:Introduction].

Winkler, W. (1990). String comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage. *Proceedings of the Section on Survey Research Methods*.

Appendix A

Similarity metrics

In Chapters 2 and 4 multiple similarity metrics were mentioned that were adapted for this research and/ or considered prior knowledge. However, for completeness these used metrics are explained with their formulas in this appendix.

A.1 Name similarities

Firstly, Jaro-Winkler similarity and Soundex encoding were used by Efremova et al. (2015), Rahmani et al. (2014) and Rahmani et al. (2016) to compare the first- and last names of people. The Jaro-Winkler similarity between two strings s_1 and s_2 is determined based on the number of matching characters m, the number of transpositions t, and an additional weight for a common prefix (Tay, 2019; Winkler, 1990). Characters are matching if these are the same and not further than $\frac{max(|s_1|,|s_2|)}{2} - 1$ positions apart. Further, the number of transpositions is half of the matching characters that are in a different sequence order. The corresponding Equation A.1 calculates the Jaro-Winkler similarity (range 0-1 with higher meaning more similar strings) with l the number of prefix characters that exactly match between s_1 and s_2 with a maximum of 4, p the weighting factor for the importance of a matching prefix (default 0.1), and Sim_J the Jaro similarity specified in Equation A.2 (Tay, 2019; Winkler, 1990).

$$Sim_{JW}(s_1, s_2) = Sim_J(s_1, s_2) + l * p * (1 - Sim_J(s_1, s_2)) \tag{A.1}$$

$$Sim_J(s_1, s_2) = \begin{cases} 0 & \text{if } m = 0 \\ \frac{1}{3}\left(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{m}\right) & \text{otherwise} \end{cases} \tag{A.2}$$

Secondly, Soundex encoding creates a string representation based on the English pronunciation of a word (Vykhovanets et al., 2020). This allows mi-

Table A.1: Soundex encoding digits with their corresponding consonants (except the letters 'Y', 'W' and 'H').

Digit	Letters
1	B, P, F, V
2	C, S, K, G, J, Q, X, Z
3	D, T
4	L
5	M, N
6	R

nor differences in spelling (due to errors or name variants) to be still mapped to the same four-character code, for example "Jansen" and "Janssen". Here, the first letter of the word is kept, then all vowels and the letters 'Y', 'W' and 'H' are removed, and consonants get represented with a number corresponding to their type of sound (see Table A.1). Then, only one occurrence of equal neighbouring digits is kept, and the code is pruned to contain only the first three digits (or extended with zeroes) (Vykhovanets et al., 2020). Thus, for instance, "Jansen" and "Janssen"would get the Soundex encoding "J525".

Thirdly, the Levenshtein distance can also be used for comparing names, which is done for place names in Sehgal et al. (2006). This distance is defined as the minimum number of character edits that are needed to be executed to convert one string into another (with a higher value meaning more dissimilar strings) (Haldar & Mukhopadhyay, 2011). The possible operations, with all their own cost (for example 1 for each) are: delete, insert, and substitute (Haldar & Mukhopadhyay, 2011).

A.2 Date similarity

Furthermore, comparing (birth, marriage or passing) dates can be done in multiple ways (Geel et al., 2012). For example, Jaccard similarity can be computed to determine how many of the components (i.e. day, month, and year values) exactly match, whereas fractional time periods return a more gradual distance value between two dates (Geel et al., 2012). This is more useful for the purpose of this research because it should be weighted more if the years match than if the day in the year corresponds. Equation A.3 shows the fractional time between dates d_1 and d_2 relative to time period tp (Geel et al., 2012).

$$FractionTime(d_1, d_2, tp) = \frac{|d_1 - d_2|}{tp} \tag{A.3}$$

where the output is forced to be a value between 0 and 1. So, when the date difference is bigger than or equal to tp, then the maximum value of one will be returned (Geel et al., 2012).

A.3 Location similarity

Moreover, comparing coordinate pairs can be done using the inverse coordinate distance from Equation A.4 as used by Sehgal et al. (2006).

$$CoordSym(l_i, l_j) = \frac{1}{dist(l_i, l_j)} \tag{A.4}$$

where $dist(l_i, l_j)$ is calculated using the Haversine distance by Balsebre et al. (2022) and Euclidean distance by Deng et al. (2019). Here, higher values mean that the locations are closer, but no maximum value exists.

In more detail, the Haversine distance (see Equation A.5) can be used to determine the distance between places on a spheroid like the Earth (Maria et al., 2020). For this, longitude-latitude coordinate pairs thus do not have to be projected onto a different coordinate reference system than WGS84 (QGIS, 2020).

$$\begin{aligned} HaversineDist(p_1, p_2) = 2r \arcsin \sqrt{\sin^2\left(\frac{lat(p_2) - lat(p_1)}{2}\right)} + \\ \cos(lat(p_1)) * \cos(lat(p_2)) * \\ sin^2\left(\frac{lon(p_2) - lon(p_1)}{2}\right) \end{aligned} \tag{A.5}$$

where p_1 and p_2 are the places of which their coordinates (longitude (lon) and latitude (lat) values in degrees) get compared (D'Agostino & Dardanoni, 2009). Further, r is the radius of the spheroid, here the Earth with $r = 6367.45$ kilometers.

Likewise, the Euclidean distance (see Equation A.6) can be used to compare the closeness of places based on the Pythagorean Theorem (D'Agostino & Dardanoni, 2009). For this it is important that both places are in the same coordinate reference system that has limited distance distortions (QGIS, 2020), and represents the places on a flat plane (Maria et al., 2020).

$$EuclideanDist(p_1, p_2) = \sqrt{(x(p_2) - x(p_1))^2 + (y(p_2) - y(p_1))^2} \tag{A.6}$$

where p_1 and p_2 are places of which their coordinates (x and y values in meter or other equal unit) get compared (D'Agostino & Dardanoni, 2009).

Appendix B

Example content similarity calculation

This appendix showcases an example calculation of identifying equivalent person entities between passport requests and candidate personal record cards. For this, the steps of Methods Sections 4.3 and 4.4 are followed on the example records shown in Table B.1.

Below insights are given into the calculations that determine each resulting feature similarity value that is shown in Table B.2.

1. First- and last name similarity calculations use the Jaro-Winkler similarity of Equation A.1, and are uniformly averaged into a mean person name similarity value. Only high values can be expected here due to corresponding blocking key strings. Further, Candidate 2 also has the same first name prefix as the passport request, which ensured an additional small similarity increase.

2. Birth date similarity calculations use the fractional time period similarity of Equation 4.2 with a time period of 10 years. Candidate 2 is just born within this threshold, so still got a similarity value higher than 0.

3. Birth place similarity calculations first check if their entity URIs correspond when records are place enriched. This is the case for Candidate 1, so directly results in the maximum place similarity value. However, for Candidate 2 the place feature similarities still need to be calculated to get a gradual place similarity value between these different place entities:

   (a) The place name similarity calculations use the normalized Levenshtein distance of Equation 4.3. Because a few letters occur in both place names, this distance is higher than 0.

Table B.1: Example fictional (translated) passport request with two candidate personal record cards to calculate similarities with. Here, the passport request and candidate 1 correspond to the records of Table 3.1, after performing pre-processing and data enrichment as discussed in Methods Section 4.1. All these records have the blocking key "mar_na_dev_es_M625_D162", which is determined as stated in Methods Section 4.2.

Features	**Passport request**	**Candidate 1**	**Candidate 2**
`FirstNames`	Maria Johanna	Maria Johanna	Mariah Joanna
`LastName`	de Vries	de Vries	de Vries
`BirthDate`	1923-04-21	1923-04-21	1932-08-30
`BirthPlace`	Jakarta	Jakarta	Garut
`BirtPlaceURI`	http://www.wikidata.org/entity/Q3630	http://www.wikidata.org/entity/Q3630	http://www.wikidata.org/entity/Q833632
`BirthPlaceLongitude`	106.84513	106.84513	107.9
`BirthPlaceLatitude`	-6.21462	-6.21462	-7.216667
`BirthCountry`	Indonesia	Indonesia	Indonesia
`BirthCountryURI`	http://www.wikidata.org/entity/Q252	http://www.wikidata.org/entity/Q252	http://www.wikidata.org/entity/Q252
`BirthProvince`	NaN	NaN	NaN

(b) The country name similarity calculations first check if their URIs correspond, which is the case here. Hence, the country name similarity gets maximum value instead of a gradual one with the normalized Levenshtein distance of Equation 4.3.

(c) The province name similarity calculations result in a missing value because no province indicators are recorded in this case. Otherwise the normalized Levenshtein distance of Equation 4.3 would have been calculated.

(d) The coordinate similarity calculations use the fractional distance similarity of Equation 4.4 with a threshold of 200 kilometers. Because the birth places are both known to be in Indonesia, the coordinates first get projected into its respective coordinate reference system EPSG:4897 to limit distance distortions. Then the Euclidean distance of Equation A.6 is used within Equation 4.4 to determine a gradual geographic distance similarity value. Because the places are 117.34 kilometers apart, this falls within the distance threshold and hence a value larger than 0 got determined.

(e) A uniform average is taken of these calculated place feature similarity values for the final place similarity, disregarding missing values (i.e. the province name similarity in this example).

4. Birth place similarity calculations are mostly only based on place name similarity calculations (as shown above for Candidate 2) when records are not place enriched. However, in a few cases both documents have recorded a country or province name. This means that for Candidate 1 also a place similarity calculation needs to be done, and that Candidate 2 gets a less informative similarity value.

5. The final similarity value is the uniform average of the mean person name, birth date, and mean place similarity values. For Candidate 1 these were all exact matches so resulted in a maximum value of one, but for Candidate 2 a more gradual value is given.

6. The candidate link is Candidate 1 because its final similarity value is larger than the one of Candidate 2. When this value is also larger than a certain threshold, then the candidate link is considered an equivalent entity.

Table B.2: Results of content similarity calculations for the records from Table B.1. The numbers in the "Similarity" column correspond to the enumeration of the calculation steps above. Here can be seen that Candidate 1 is a perfect match for the passport request, while Candidate 2 is likely not the same entity but still has the same blocking key. If multiple numbers are mentioned, separated by a semicolon, then the first and second numbers respectively represent the similarity score from non- and place enriched data. If only one number is mentioned, then the value is the same in both situations.

Similarity	**Candidate 1**	**Candidate 2**
1. First name	1.00	0.97
1. Last name	1.00	1.00
1. **Mean person name**	1.00	0.99
2. **Birth date**	1.00	0.06
3a. Place name	1.0; NaN	0.29
3b. Country name	NaN	NaN; 1.00
3c. Province name	NaN	NaN
3d. Coordinates	NaN	NaN; 0.41
3(e). **Mean place**	1.00	0.29; 0.57
5. **Mean place, birth date, and person name**	1.00	0.45; 0.54

Appendix C

Context similarity calculation

Next to the content similarity of Section 4.3, computing a context similarity between two documents can increase their total similarity value when mentioned reference(s) in one document corresponded to reference(s) in the other one. How to apply this procedure is explained in this appendix to be able to apply it when all meta data of the personal record cards in this case study is complete, to research the influence of place enrichment on another promising baseline method.

The set of references that can be used for computing the context similarity consist of the people from the personal record cards and passport requests that were not used to compute the content similarity between, for example partner(s), child(ren), and/or parent(s) of the key person. The procedure and equations for this calculation can be a combination of the methods designed by Rahmani et al. (2014) and Rahmani et al. (2016). As proposed combination the context similarity of Rahmani et al. (2014) is computed between documents linked by matching blocking keys of recorded relatives and multiplied by the probability to reach that document with a random walk within a certain number of document links, like Rahmani et al. (2016) does. This way, both second-level (family) relationships are considered as well as the proportion of documents with the same blocking key.

This combined context similarity value is thus determined proportionally to how likely it is that two references (r_i and r_j) occur in the same block b_k due to having the same blocking key string (Rahmani et al., 2014). This is expressed as the confidence $Conf(b_k)$ in Equation C.1. Here, a bigger number of references belonging to a block (i.e. larger $size(b_k)$) decreases the confidence that these two references are identical (Rahmani et al., 2014).

$$Conf(b_k) = \frac{N}{size(b_k)} - 1 \tag{C.1}$$

where N is the number of all references in all collections.

Furthermore, Equation C.2 returns the blocking context of certificate c_i, which are the block ids of the people referenced in this document (Rahmani et al., 2014).

$$BC(c_i) = \bigcup_{r_j \in c_i, r_j \in b_k} b_k \tag{C.2}$$

Moreover, in Equation C.3 the outputs of Equations C.1 and C.2 are then used to calculate a final context similarity value between the key persons r_i and r_j of respectively certificates c_i and c_j (Rahmani et al., 2014). This is calculated as a weighted proportion of the number of equivalent references with respect to all their mentioned references (Rahmani et al., 2014).

$$Sim_{BC}(r_i, r_j) = \frac{\sum_{b_k \in \{BC(c_i) \cap BC(c_j)\}} Conf(b_k)}{\sum_{b_k \in \{BC(c_i) \cup BC(c_j)\}} Conf(b_k)} \tag{C.3}$$

Alternatively, the content similarity equations could be utilized for each other reference mentioned in a document. This would probably (slightly) further increase the likelihood that correct document links are made that refer to equivalent person entities, but at the expense of higher computational power. Because many genealogical documents exist in the data set of this case study, especially when in the future other collections are being connected as well, it is infeasible to perform these additional calculations while these described context similarity calculations likely already achieve a low number of false positives.

Finally, the addition of content and context similarity gives a similarity value between 0 and 2, as in Rahmani et al. (2014) and Rahmani et al. (2016), with a higher value meaning a higher confidence that the key people of the documents refer to the same entities.

Appendix D

Converting entity resolution output into linked open data

After the entity resolution process, the outputs can be converted into a LOD format to enrich the resulting database as proposed in Section 6.3. This appendix shows an example of this procedure which can be applied after future research defined a suited LOD ontology for CBG.

To apply linked open data principles to genealogy, each document in a collection and its referenced entities first need to get (HHTP) URIs instead of raw strings (Cuijuan et al., 2018). URIs of already recorded entities can be found via LOD sources such as Wikidata, and new ones can be added as well. Next, their relationships with other entities should be described in a standardized way using a suited predicate from the defined ontology (Cuijuan et al., 2018). With such entity and relationship URIs, more information can be looked up that is not recorded within a document, for example about a referenced place, person, or time period (Berners-Lee, 2009; Blaney, 2021).

Examples of existing ontologies that describe relationships or contain information about unique entities are: schema.org and WikiData for many different purposes, GeoNames for places, LOC'S BIBFRAME 2.0 for bibliography control, Wide Web Consortium (W3C)'s Time Ontology for temporal periods (Cuijuan et al., 2018), Historical International Standard Classification of Occupations for historical jobs (Meroño-Peñuela et al., 2016), and FOAF and Bio CRM for information about people (Leskinen & Hyvönen, 2020; Leskinen et al., 2017). Because Dutch municipalities changed a lot through history, additional links to gemeentegeschiedenis.nl can be created (which refers in turn to GeoNames and DBpedia) (Meroño-Peñuela et al., 2016).

In more detail, first the key person of each personal record card has to be associated with a unique person entity, similarly to the procedure by Koho et al. (2020). Then, each passport request that can not get linked with a personal record card also needs to have their key person be associated with

a new person entity (Koho et al., 2020). The FOAF ontology of `Person` could for example be used for representing person entities (Leskinen et al., 2017), but future research has to first identify the best suited ontology for CBG. Hence, this example uses a unique *ID* for each found person entity and a corresponding placeholder URI: *http://cbg.nl/person/ID*.

Second, a triple in `Turtle` format should be created to link these person entities to the personal record cards and/or passport requests that the entity resolution procedure found to refer to them. For example, the `relatedLink` predicate of schema.org can be used for this, which has the URI https://schema.org/relatedLink (Leskinen & Hyvönen, 2020). Alternatively, the `publications` predicate of FOAF can be used for the same purpose, but that is better suited for published articles by this person (Kassiri & Belouadha, 2017).

Third, all of the referenced documents are expected to be in a website URL format by the `relatedLink` predicate (Schema.org, 2022), but can in the future also get an URI like *http://cbg.nl/person_card/ID*. These document entities can then, for example, use the FOAF ontology of `Document`. This way more meta data about this document, like an URI to the GeoNames entity of a referenced place (Meroño-Peñuela et al., 2016) or information about its storage location, can also be represented in a standardised LOD structure.

As a result of these steps, an example of a `Turtle` triple (with placeholders) that future research can represent the entity resolution output in can thus be:

<http://cbg.nl/person/000001>
<https://schema.org/relatedLink>
<http://cbg.nl/person_card/000001>

Lastly, remaining features referenced within the linked documents can also be converted into LOD and aggregated to the person and document entities (after potential conflict resolution (Knap et al., 2012)) (Koho et al., 2020). Specifically for the retrieved GeoNames and/or WikiData place entities in this research, the `birthPlace` predicate of schema.org can be used to describe their relationship with person entities, which has the URI https://schema.org/birthPlace (Nurmikko-Fuller et al., 2016; Pellissier Tanon et al., 2020). With such a LOD relationship instead of raw strings of features, more information about that place can be retrieved as well as information about other people that are related to that place. An example `Turtle` triple (with placeholders) for a relationship with the Dutch city Utrecht can thus be:

<http://cbg.nl/person_card/000001>
<https://schema.org/birthPlace>
<https://www.sws.geonames.org/2745912/>