Small Lives, Big Meanings
Expanding the Scope of Biographical Data through Entity Linkage and Disambiguation

Lodewijk Petram, Jelle van Lottum, Rutger van Koert, Sebastiaan Derks
Huygens ING, KNAW Humanities Cluster
Oudezijds Achterburgwal 185, 1012 DK Amsterdam
E-mail: {lodewijk.petram; jelle.van.lottum; sebastiaan.derks}@huygens.knaw.nl; rutger.van.koert@di.huc.knaw.nl

Abstract

The Huygens institute for Dutch history and culture aims to facilitate and enhance collaborative research with and on biographical data. We give a brief outline of the Huygens ING digital biographical data policy, describe how we share our data with the world, and explain how we facilitate the exploration of similarities and interconnections between the Huygens data, external data collections and user-uploaded datasets, without imposing selection criteria. Finally, we present a use case that shows how our policy and infrastructure enable researchers to employ large collections of ambiguous biographical data, hitherto mainly used for genealogical reference, for addressing innovative, challenging research questions.

Keywords: biographical data, entity matching, disambiguation, digital infrastructure, genealogical data, prosopography

1. Introduction

Daniel Engel was born in Danzig (present-day Gdańsk in Poland) and signed up with the Delft branch of the Dutch East India Company (VOC) on the first of October, 1766. He worked as an ordinary seaman during the seven-month journey to Batavia (now Jakarta), stayed there for nine months and then sailed back to Europe on the same ship. Engel was probably illiterate, as he signed with a cross.[1]

This is all we can infer about the life of Daniel Engel from his employment record – too little by far to deserve an entry in the Biography Portal of the Netherlands[2], the online collection of biographies of prominent people from Dutch history, maintained by the Huygens Institute for Dutch history and culture (Huygens ING). Engel was simply one of the many thousands of men from German lands who joined the ranks of the VOC in the seventeenth and eighteenth centuries.

But there is more on Daniel Engel. It seems he joined the VOC two more times, in 1788 and 1792, as a boatswain's mate and able seaman, respectively. The latter employment record furthermore shows that Engel died in Asia, on the second of October, 1798. There is also mention of a Daniel Engel from Danzig in the interrogation transcripts of the English admiralty, dating from the Fourth Anglo-Dutch War (1780-1784), when the English seized many Dutch ships. This sailor worked as a boatswain on a merchant's ship that was supposed to have brought cargo from Curacao to Rotterdam in 1782. He was born in 1753 or 1754.[3]

It is likely that these four data observations refer to the same individual: together they form a logical career path of an eighteenth-century sailor, even though Daniel Engel would have been only twelve or thirteen years old when he first sailed to Asia. This mini-biography is hardly revolutionary – historians have pieced together bits of biographical information from multiple sources for ages (e.g. Ogborne, 2008) – and it is also still not worthy of an entry in the Biography Portal. However, advances in digital techniques now allow for (semi-)automated matching of large numbers of data entities. Disambiguating the just under 800,000 person entities in the VOC employment records has become feasible, and the same holds for data observations in other large, digitized source collections. This opens up possibilities for employing the many snapshots of persons' lives that are available in e.g. genealogical sources and historical employment records in large-scale prosopographical analyses that may be instrumental in answering urgent, challenging research questions.

At Huygens ING, we seek to connect such collections of disambiguated data to our traditional, mostly highly curated sets of biographical data, with the intent to create an integrated environment that meets the needs of researchers working on a broad range of research questions. In the remainder of this paper, we outline the Huygens ING digital biographical data policy and how we aim to incorporate data on the lives of both prominent people and small fry in our new linked open data infrastructure, give a short overview of the technique we use for (semi-) automatically matching entities from one or multiple sources, and finally present a research use case.

2. Huygens ING and Biographical Data

The mission statement of Huygens ING reads: 'Innovating history: unravelling history with new technology'. The institute tries to accomplish this mission by developing and applying new, advanced digital tools that help open up

[1] These and other VOC employment data: VOC Opvarenden database (http://www.gahetna.nl/collectie/index/nt00444/view/NT00444_OPVARENDEN and http://dutchshipsandsailors.nl/).

[2] http://www.biografischportaal.nl/

[3] Prize Paper Dataset, cf. footnote 6.

historical sources, which are often difficult to access and use, and hence stimulate innovation in research. The institute's updated digital biographical data policy reflects this mission.

Traditionally, biographical dictionaries have formed the heart of the Huygens ING biographical data collection. The institute has a long history of editing biographical dictionaries and publishing these as book series or, in more recent times, making the entries available digitally through separate web interfaces. The development of the Biography Portal, essentially an index to the various biographical dictionaries, was a first effort of bringing together the available biographical data.

Huygens ING is now gradually entering a new stage, in which all biographical data are migrated to the institute's new digital infrastructure. Structured data on person entities are interlinked with a text browser, in which the original texts of the biographical dictionaries and other book and source collections are made available. A user can thus easily search for a person and view related entries in biographical dictionaries and mentionings in other texts.

So far, the new infrastructure largely resembles a re-fashioned Biographical Portal. What is new, however, is that the structured data are ingested into a linked open data (LOD) environment, and can hence easily be linked with other datasets (both internal and external, national and international). To guarantee optimal findability and re-usability of our data on persons, we align all person entities to those linked open data ontologies that are most used in the Arts and Humanities, and by cultural heritage institutions, both within the Netherlands and internationally: CIDOC-CRM, Wikidata, schema.org and FOAF.

Furthermore, the new digital infrastructure is specifically designed as a humanities research environment. Whereas the traditional book volumes, web interfaces, and even the Biography Portal first and foremost served as reference works – a typical researcher would use them to look up information on one or a small number of persons – the new environment offers better search (elasticsearch) and functionality to explore similarities and interconnections, thus allowing users who practice collective biography and prosopography to easily collect data on the groups of people of their interest (cf. Harders and Lipphardt, 2006). Researchers can furthermore link data elements across multiple sources and use data observations to enrich their own datasets. Finally, they can query the data through the API or download selections of data in various file formats, and then analyse the data offline or using tools for data analysis and visualisation that are available on the internet. In short, the data are ready to be used by researchers.

The Huygens ING digital infrastructure thus has an interactive character; the focus is not solely on making data available, but also, and especially, on allowing researchers to use and share them. In parallel to this, we aim to facilitate and enhance collaborative research with and on biographical data, which comprises, in our view, any biographical data that might be of interest to academia. Researchers are welcome to upload their own data, which they can link up to our data, or make connections between the data in the infrastructure and external datasets. Furthermore, to accommodate researchers' needs, we are currently developing an entity matching tool, which will become available within the digital infrastructure, that allows researchers to easily find candidates of matches between entities from multiple datasets. After validation, the matches will be linked to a resolved entity. We will go further into the details of this tool in the next section.

To gather and present the data in clear, domain-specific collections, our infrastructure consists of multiple, interconnected instances. The curated Huygens ING datasets on the history of knowledge, Dutch history and literary studies are available for reference and analysis in Data Huygens ING.[4] This data hub is directly linked to that of CLARIAH[5], the Dutch national digital infrastructure project for the Arts and Humanities. The benefit of this set-up is that it enables us to validate and manage the data within the domain context, and it also helps us implement our data provenance policy. Huygens ING provides comprehensive provenance information for all its datasets and presents this in a form that is both understandable for humans and interoperable with other data infrastructures within the semantic web. On dataset level, the provenance information consists of a short and general description of the dataset, a list of most-used sources, and information on selection criteria and information extraction techniques that were applied in the process of compiling the dataset. This information is available as an introductory text to the dataset and is also added, in short form and modelled using the P-PLAN Ontology (Garijo and Gil, 2012), to every record. As such it will enable researchers who see an isolated data observation in the LOD cloud to learn about the context in which the data observation came about. Additionally, on record level, we provide specific references to sources. We encourage users to provide the same information for user-uploaded datasets in the CLARIAH data hub. Furthermore, for all data in the infrastructure, technical provenance information is automatically retained. This allows users to see when a particular dataset was originally uploaded and by whom, and which edits were made on a particular data element, either manually or by built-in tooling, such as for entity matching.

3. Automated Record Linkage

The record linkage tool we are currently developing enables users to find matches between entities in one or more sets of data observations, selected from the structured data repository within our digital research environment or external LOD sources. For the time being, the tool is primarily intended for finding matches between person entities. It allows users to measure name similarity and refine candidate matches using rules that are e.g. based on geographical data or dates.

We chose to develop the tool in a PostgreSQL environment for the relatively speedy matching results it offers, especially when using trigram matching. The tool downloads selected rdf triples, automatically converts them

[4] https://data.huygens.knaw.nl/

[5] https://www.clariah.nl/; https://anansi.clariah.nl/

into csv-format and loads them into the PostgreSQL environment. In the matching process, it creates a new dataset with matched entities, which, after validation by the user, is returned to the LOD environment. This new dataset includes full provenance information about the matching parameters that were applied (algorithm and additional rules) and the user doing the final validation step. All provenance data are automatically retained during the process of candidate generation and validation.

The tool offers various methods for measuring string similarity, which can be used for matching names and toponyms: trigram matching (the preferred method, for speed reasons; it uses the similarity function in the PostgreSQL (9.5) pg_trgm module), Levenshtein distance, and (Double) Metaphone. When geocodes are available, locations can also be matched using the PostgreSQL extension PostGIS. This extension allows users to find matches based on either an exact geographical location or a user-set range around a geographic point.

To start the matching procedure, a user first manually selects data fields for matching and then creates a set of refinement rules, tailored to the data at hand, to improve matching results and/or exclude irrelevant matching candidates. For example, if a user wants to match entities from a birth register with a faculty list, he could create a rule that discards candidates that would have been under eighteen or over one hundred years of age when employed at university. Another rule could state that candidates who are between age 25 and 65 when employed at university should get higher scores.

The tool leads the user through an iterative matching procedure (cf. e.g. Efremova et al., 2014; Idrissou et al., 2017). Users are encouraged to set strict rules at first. This will yield a relatively small number of high-quality candidates, from which the user can then select matches for approval. After this first matching and validation round, the matched records are split from the original dataset and sent to a new dataset, which only contains validated data. The user can then let the tool iterate once or multiple times over the remaining original data using different sets of matching rules to generate additional candidate sets, from which approved matches can be added to the set of validated data. Taken together, the steps in the matching procedure yield results with high precision and recall.

4. Research Use Case: Sailors' Careers

The research project 'Human capital, immigration and the early modern Dutch economy: job mobility of native and immigrant workers in the maritime labour market, c.1700-1800 (HUMIGEC)' illustrates the potential of our infrastructure and entity linkage tool for academic research. This project's research question originates from a currently hotly debated topic in both the political and the public arena: what is the economic contribution of migrant workers on a recipient economy? This is a difficult question to answer for modern economies, let alone for economies from the past, since historical statistics on education or training levels of workers are largely lacking. Without these, simply having estimates of the size of the migrant influx is not sufficient. After all, it makes a huge difference whether migrants are non-skilled, skilled or become skilled during their careers in the recipient economy.

Although for the pre-1800 period sources containing clear indicators of education or training levels are rare, we do have large numbers of historical employment records. However, such sources often provide no more than a snapshot of a person's life and are therefore relatively limited in their use. But by matching entities from multiple source collections, these records become much more meaningful. Matching a sufficiently large number of entities was hitherto practically impossible, due to the simple fact that these data collections are large and manually finding matches takes a lot of time, but our automated entity matching tool enables us to do so – and in the near future other scholars as well. In the case of HUMIGEC, the tool helps us to reconstruct individual careers, which in turn makes it possible to compare the relative successfulness of migrant and native workers. As the success of careers is a good indicator of skills, this assessment will allow us to address the central research question of the project.

We selected the maritime sector of the eighteenth-century Dutch Republic as a case study in HUMIGEC, because this was a key sector of the economy, characterised by a high level of migrant participation. Moreover, its workers were well documented: we have almost 800,000 employment records of the VOC, digitised by a number of archival institutions in the Netherlands, that cover the entire eighteenth century, and c. 15,500 records on Dutch mercantile marine crews from the Prize Paper Dataset compiled by HUMIGEC's PI Jelle van Lottum.[6] Each record in both collections contains data on a sailor's name, place of birth, rank on board and start date of the employment. For the sailors in the Prize Paper Dataset, we also know their age when questioned by the English admiralty.

By matching entities within and between these datasets, as shown by the example of Daniel Engel in the introduction to this paper, we can (partially) reconstruct sailors' careers, which we can then use to compare the level of job mobility (i.e. promotion or job switching) of non-migrant and migrant workers (Gibbons and Waldman, 1999). This will give us insight into the extent to which migrants succeeded in gaining skills (i.e. human capital) during their careers, and compare this to non-migrants.

We use the entity alignment tool introduced in the previous section to find data observations that are probably related to the same individual. We first look for data observations with a high level of name similarity, measured on the basis of trigram matching, and filter out irrelevant results by applying a set of rules based on dates (for example, a person cannot have sailed out before birth or after death, cannot have been employed on two ships at the same time, cannot have been in Asia and Europe at the same time, etc.) and domain expertise (it is e.g. unlikely that a person who had worked as an ordinary seaman on a

[6] Data collected in the Economic and Social Research Council (ESRC) project no. RES-062-23-3339: Migration, human capital and labour productivity: The international maritime labour market in Europe, c. 1650–1815.

single trip rejoined the ranks of the VOC as a captain).

Next, we use the sailors' places of birth as a check on matches. Since we have to deal with quite a bit of variation in toponym spelling – all records were written down by clerks who often did not speak the same language as the sailors in front of them and who were also frequently unfamiliar with the towns and villages, often in the German lands and Scandinavia, mentioned by the sailors – we decided to try to standardise place names and reconcile them to their modern-day GeoNames equivalents. We have so far standardised around 30,000 unique toponym attestations and aim to at least double this number before project end. The standardised toponyms allow us to first look for exact matches. Thereafter, using the geo coordinates given back by GeoNames, we geo-group locations to find possible additional matches. In this way, we also catch sailors who used their birth place and region interchangeably.

For all remaining person entity matches, suggested on the basis of name similarity, but not corroborated by matching places of birth, we perform a final birthplace check by measuring string similarity of the original place name attestations, so as to account for possible mistakes by clerks or transcribers of the original documents – Norden in East Frisia might easily have been misunderstood as Naarden close to Amsterdam.

The scope of our project does not allow for experiments with standardising person names. We therefore rely on the trigram matching algorithm to cope with spelling variations in names. However, for a follow-up project to HUMIGEC, we are thinking of also standardising person names, beginning with native workers' names. To this end, we would use the Database of Surnames in The Netherlands[7] to standardise family names, and group variants of given names on the basis of data generated by Gerrit Bloothooft (e.g. Bloothooft and Schraagen, 2015).

This paper is not well-suited for going deeply into socioeconomic analysis and statistical results – incidentally, HUMIGEC is still an ongoing research project and we currently only have very preliminary results – but a brief reflection on methodology is in place. First of all, it is important to stress that our method is far from perfect. At best, it gives us a limited view on career paths in the Dutch eighteenth-century maritime sector, for the available sources do not cover the entire sector and we have no ground truth for assessing the performance of the entity linkage process. We do, however, have a set of manually-matched entities that we use for a superficial assessment of our matching method. However, these matches are self-evidently incomplete and are furthermore likely to be biased towards non-standard names.

That same bias will be present in the automatically generated matching candidates: disambiguating employment records of sailors with common names, who were born in large towns and cities, is in many cases simply impossible, both for humans and computers. This gives reason for some concern about the representativeness of our study, but then again, sailors with non-standard names were not atypical because of their unusual names. Sixtus van den Hoek from Delft, for example, a sailor we could easily trace in eight different VOC employment records, is not unrepresentative because of his name – were his name Jan de Jong, he would not suddenly become a synecdoche for maritime life (cf. Van Lottum, Brock and Sumnall, 2015). Moreover, since we base our analysis on a large number of observations, we think the bias in our sample towards non-standard names will not have a significant influence on our results. However, to check whether the non-standard names that are likely to be overrepresented in our sample were not typical for a certain class of eighteenth-century society, we will compare them to the family names of Amsterdam's highest-income tertile, derived from registers of a 1742 income tax (Oldewelt, 1945), and to the family names in Amsterdam's birth, marriage and death registers from the mid-eighteenth century.[8]

A discussion of the digital heuristics involved in our project will naturally also be included in the general description of the set of validated record matches and, in very short form, in each record's P-Plan provenance. So, if for example a future researcher of the Asian activities of the VOC would see that some person entities from the Official letters of the United East India Company – a Huygens ING digital resource that will be added to our LOD infrastructure in due course – were connected to records detailing sailors' careers and others not, he would know that this could have as much to do with selection bias in the linkset as with the actual careers of these people.

5. Conclusions

Biography as a historical method has traditionally mainly been used as a means to illustrate qualitative themes, generally based on one or a small set of case studies. From around the turn of the century the online availability of national biographical dictionaries in e.g. the Netherlands, Germany, the United Kingdom and Australia allowed for larger-scale biographical research and the formation of collective biographies (cf. Arthur, 2015; Carter, 2012). But these were inevitably limited by the scope of the online biographical collection and influenced by the selection criteria (and biases) of its editors.

The Huygens research infrastructure and biographical data policy, however, allow researchers to go one step further. The institute makes available all biographical data contained in its collection, both highly curated data from biographical dictionaries and persons data retrieved from various textual sources. Furthermore, as illustrated by the HUMIGEC research case, researchers can use the infrastructure to semi-automatically connect external datasets to the core data or disambiguate their own data. In HUMIGEC, we use the large number of mini-biographies obtained through digital methods as a means of illustrating wider social and economic processes. Indeed, as Paul Arthur predicted, this approach is 'a demonstration of biography's greatly increased capacity, in the digital era, to activate cross-disciplinary investigation, and become a dynamic agent for integrating and connecting individual lives and their historical contexts' (Arthur, 2015).

[7] http://www.cbgfamilienamen.nl/

[8] https://archief.amsterdam/indexen/

Digital advances such as the one described in this paper are blurring the boundaries between (collective) biography, prosopography and other socioeconomic research methods. In parallel with this development, all biographical data observations, however insignificant they may seem at first sight, may become very meaningful and instrumental to answering important research questions when disambiguated and combined with other data. Huygens ING aims to facilitate and enhance the full range of biography methods by making available a digital infrastructure that welcomes all biographical data – be they on the lives of prominent people or small fry – and offering functionality for exploration of similarities and interconnections between data observations.

6. Acknowledgements

We thank Ania Ahamed and Jessica den Oudsten for their research assistance, and the Amsterdam City Archives for sharing their genealogical data with us. HUMIGEC received funding from CLARIAH.[9]

7. References

Arthur, P. (2015). Re-imagining a Nation: The Australian Dictionary of Biography Online. *European Journal of Life Writing*, 4, pp. 108--124.

Bloothooft, G., Schraagen, M. (2015). Learning Name Variants from Inexact High-Confidence Matches. In G. Bloothooft, P. Christen, K. Mandemakers, M. Schraagen (Eds.), *Population Reconstruction*. Cham: Springer, pp. 61--83.

Carter, P. (2012). Opportunities for National Biography Online: The Oxford Dictionary of National Biography, 2005–2012. In M. Nolan, C. Fernon, *The ADB's Story*. Canberra: ANU Press, pp. 345--371

Efremova, J., Ranjbar-Sahraei, B., Oliehoek, F.A., Calders, T., Tuyls, K., (2014). A Baseline Method for Genealogical Entity Resolution. *Proceedings Workshop Population Reconstruction: 19-21 February 2014, Amsterdam*.

Garijo, D., Gil, Y. (2012). Augmenting PROV with Plans in P-PLAN: Scientific Processes as Linked Data. *Proceedings of the 2nd International Workshop on Linked Science: 12/11/2012, Boston, USA.*

Gibbons, R., Waldman, M. (1999). Careers in organizations: theory and evidence, in: Ashenfelter, O., Card, D. (Eds.), *Handbook of labor economics*. Vol. 3B. Amsterdam: Elsevier, pp. 2373--2437.

Harders, L., Lipphardt, V. (2006). Kollektivbiografie in der Wissenschaftsgeschichte als qualitative und problemorientierte Methode. *Traverse*, 13(2), pp. 81--91.

Idrissou, A.K., Hoekstra, R., Harmelen, F. van, Khalili, A., Besselaar, P. van den (2017). Is my sameAs the same as your sameAs? Lenticular Lenses for Context-Specific Identity. *Proceedings of the Knowledge Capture Conference, Austin, TX, USA, December 04 - 06, 2017.* Article No. 23.

Lottum, J. van, Brock, A., Sumnall, C. (2015). Mobility, Migration and Human Capital in the Long Eighteenth Century: The Life of Joseph Anton Ponsaing. In: M. Fusaro et al. (Eds.), *Law, Labour, and Empire. Comparative Perspectives on Seafarers, c. 1500-1800*. Basingstoke: Palgrave Macmillan, pp. 158--176.

Ogborne, M. (2008). *Global lives. Britain and the world, 1550-1800*. Cambridge: Cambridge University Press.

Oldewelt, W.F.H. (1945). *Kohier van de personeele quotisatie te Amsterdam over het jaar 1742.* 2 vols. Amsterdam: Genootschap Amstelodamum.

[9] https://www.clariah.nl/en/projects/research-pilots/granted-pilot-research-projects/humigec