Amsterdamers from the Golden Age to the Information Age via Lenticular Lenses

Al Idrissou[ab] Veruska Zamborlini[c] Chiara Latronico[c] Frank van Harmelen[a]
Charles van den Heuvel[d]

[a]*Department of Organization Sciences, Vrije Universiteit Amsterdam*
[b]*Dept. of Computer Science, Vrije Universiteit Amsterdam, The Netherlands*
{o.a.k.idrissou, frank.van.harmelen}@vu.nl
[c] *Institute for Logic, Language and Computation, University of Amsterdam, The Netherlands*
{v.zamborlini, c.latronico}@uva.nl
[d] *Huygens Institute for the History of the Netherlands*
charles.van.den.heuvel@huygens.knaw.nl

1 Introduction

The Golden Agents infrastructure project[1] closely collaborates with the Amsterdam City Archives (SAA) to publish their digitized registries as Linked Open Data (LOD). In their All Amsterdam Acts project[2], the SAA digitizes/indexes all its notarial acts. For Golden Agents, studying the interactions between the production and consumption of the creative industries of the Dutch Golden Age is relevant because the probate inventories, testaments, etc. in these acts reveal the objects that families living in Amsterdam had in their houses. However, to link these data to other relevant collections we need to disambiguate names to identify individuals. This is a challenging task because (i) citizens in the Dutch Golden Age were not given any identification number; (ii) the information supplied in a single index do not suffice to uniquely identify an individual (weak identity criteria) and (iii) because of multiple occurrences of a single individual within an index. Here we discuss the requirements we identified for addressing this challenge (Sect. 2); the Lenticular Lenses as an innovative context-sensitive entity linking method (Idrissou et al., 2017) (Sect. 3) and our first experiments applying this tool for connecting three SAA indexes (marriage, baptism and probate inventories) and two authoritative datasets (Ecartico[3] and ULAN[4]) (Sect. 4).

2 Requirements

This section describes and exemplifies five requirements using our case study:

1. Identify identity criteria properties:
   - Full names and dates;
2. Identify matching methods:
   - ***String approximation***: to account for small name variations;
   - ***String mapping via authoritative datasets***: to account for big variations;
   - ***Date approximation***: to compare dates from different registries;

[1]Golden Agents `https://www.goldenagents.org`: Creative Industries and the Making of the Dutch Golden Age is funded by the NWO Large-investments program.

[2]`http://alleamsterdamseakten.nl`

[3]`http://www.vondel.humanities.uva.nl/ecartico/`

[4]`http://www.getty.edu/research/tools/vocabularies/ulan/`

3. Generate candidate links:
   • The methods above are combined to generate candidates;

4. Analyze links:

   Here, a new iteration may start from (1), in case of data issues or if the inclusion of new properties is needed, or from (2), if new methods are needed.
   • ***Clustering*** is used to check the quality of the discovered links.

   ◦ ***New properties or methods.*** The initially selected properties has shown not to be good enough identity criteria. To strengthen it, we tested co-occurrence of pairs of resources (**cross-match**) in different records (e.g. *if a pair of groom and bride in one record appears, for example, in another record as groom and previous wife, they are more likely to be the same*).

   ◇ ***Data issues***: This process revealed that data issues may influence the quality of the links: duplication of registries weakens the co-occurrence method (e.g. *sometimes the same marriage is registered twice*); and incompatible co-occurrences (e.g. *names that appear as both groom and bride*);

5. Validation strategies (yet to be performed by domain experts)

   Batch validation: automated quality verification can allow for reducing the amount of clusters to be manually verified.

3 Lenticular Lenses Method

Lenticular Lenses (LL) is a context-sensitive alignment method which keeps track of provenance. It supports the use of several combinations of properties as identity criteria depending on the context and the use and combination of several alignment methods to generate candidate links (requirements 1-3). Furthermore, it allows for two types of analysis (requirement 4): (i) clustering alignments in a way that supports verifying the occurrence of an entity in several datasets, and (ii) creating views (tables) on the datasets based on automated SPARQL queries that can be useful for users as a starting point for more complex queries. The validation (requirement 5) is currently supported in a fine-grained level, i.e. each pair of linked resources can be validated individually. A more advanced validation is under development. This will allow for manually validating a whole cluster at once or automatically validating several clusters based on metrics.

4 Experiment

To disambiguate Amsterdamers mentioned in SAA, a strategy (see Figure 1) is to match them against an authoritative dataset such as **Ecartico** (∼30K disambiguated (alternative) names, including marriage, birth and relatives). While the SAA **marriage registries** (∼1.3MM names) describe events by registration date, groom, bride, previous husband or wife, etc. (Fig. 2); **baptism registries** (∼3.7MM names) use registration date, child, parents, etc. (Fig. 3); and **probate inventories** (∼23K names) mention registration date, people involved, etc. (Fig. 4).

First a **set of candidate matches** is created by interlinking names whenever they are **approximately similar by at least 90%** or **the same according to ULAN**. Then, we filter the matches by considering appropriate **date approximations**. Finally, we perform a **cross-match** to find pairs of names (names connected within a dataset) that are matched with other pairs.

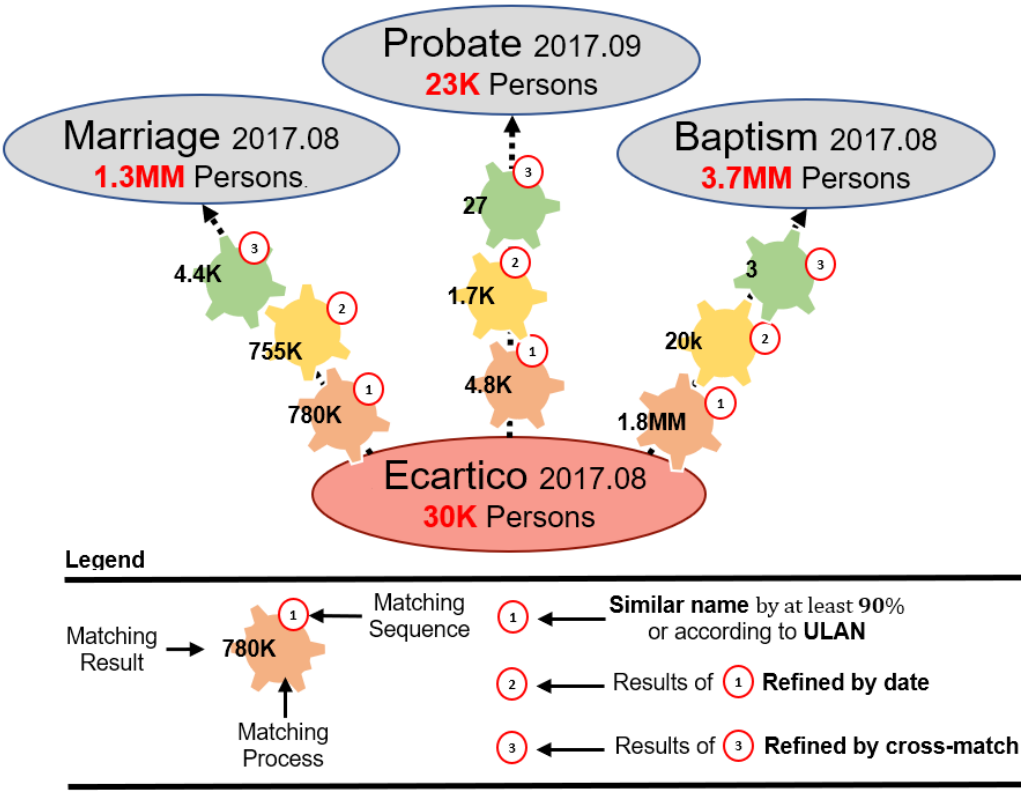

Figure 1: **Experiments Overview** - Disambiguating SAA indexes with Ecartico.

Stricter matching conditions produce less matches as shown in Figure 1[5]. Presumably, this is due to less false positives and more false negatives (to be confirmed after validation).

Naturally this experiment allows for disambiguating only the Amsterdamers mentioned in Ecartico. We consider the LL method to be promising as (i) partial verification of the final results revealed meaningful matches and (ii) new matching processes, such as transitivity, can be combined to produce better results.

[5]Data issues (e.g. missing child's surname in Baptism) need to be addressed.

```
<?xml version="1.0" encoding="UTF-8"?>
<indexRecords name="SAA Index op ondertrouwregister">
    <indexRecord id="saaId26036853">
        <Inschrijvingsdatum>1705-05-08</Inschrijvingsdatum>
        <Naam_bruidegom>Boom, Hendrik [van den]</Naam_bruidegom>
        <Naam_eerdere_vrouw>Mijnderts, Elsje</Naam_eerdere_vrouw>
        <Naam_bruid>Willems, Margrita</Naam_bruid>
        <Naam_eerdere_man>Barentsz, Gerrit</Naam_eerdere_man>
        <Bronverwijzing>DTB 704, p.333</Bronverwijzing>
        <Opmerkingen>Huwelijksintekeningen van de PUI. Akte doorgehaald, zie
            bladzijde 344</Opmerkingen>
        <urlScan>https://archief.amsterdam/inventarissen/inventaris/5001.nl.html#A25472000169.JPG
        </urlScan>
    </indexRecord>
</indexRecords>
```

Figure 2: Marriage registry

```xml
<?xml version="1.0" encoding="UTF-8"?>
<indexRecords name="SAA Index op doopregister">
    <indexRecord id="saaId23491747">
        <Kind>..., Corneelis</Kind>
        <Geboortedatum>1810-12-24</Geboortedatum>
        <Doopdatum>1811-01-06</Doopdatum>
        <Kerk>Oude Kerk</Kerk>
        <Godsdienst>Hervormd</Godsdienst>
        <Vader>Kist, Cornelis</Vader>
        <Moeder>Klapwijk, Maria</Moeder>
        <Getuige>Klapwijk, Ari</Getuige>
        <Getuige>Scholte, Elizabeth</Getuige>
        <Bronverwijzing>DTB 37, p.1(folio 1), nr.2</Bronverwijzing>
        <urlScan>https://archief.amsterdam/inventarissen/inventaris/5001.nl.html#000000040317.JPG
        </urlScan>
    </indexRecord>
</indexRecords>
```

Figure 3: Baptism registry

```xml
<?xml version="1.0" encoding="UTF-8"?>
<sets>
    <set>
        <uuid>5a49cdba-5337-a13e-42a3-f107ac9f9281</uuid>
        <notaris>DANIEL BREDAN</notaris>
        <inventarisNr>941</inventarisNr>
        <akteNr>79607</akteNr>
        <akteType>Boedelinventaris</akteType>
        <datering>1631-08-27</datering>
        <taal>nederlands</taal>
        <beschrijving>
            Boedel van de waard in de Blancken Ham. Zeven schilderijen, waaronder 4
            portretten. Stadsgezicht en de 10 geboden. Een uithangbord. Bedden,
            keukengerei.
        </beschrijving>
        <persoonsnamen>
            <persoonsnaam id="p1">
                <voornaam>Jacob Jacobsz</voornaam>
                <tussenvoegsel>van</tussenvoegsel>
                <achternaam>Schagen</achternaam>
            </persoonsnaam>
            <persoonsnaam id="p2">
                <voornaam>Vincent</voornaam>
                <achternaam>Jacobsse</achternaam>
            </persoonsnaam>
            <persoonsnaam id="p3">
                <voornaam>Gerrit</voornaam>
                <achternaam>Gerritsen</achternaam>
            </persoonsnaam>
            <persoonsnaam id="p4">
                <voornaam>Trijn</voornaam>
                <achternaam>Jans</achternaam>
            </persoonsnaam>
            <persoonsnaam id="p5">
                <voornaam>Hans</voornaam>
                <tussenvoegsel>ter</tussenvoegsel>
                <achternaam>Kinderen</achternaam>
            </persoonsnaam>
        </persoonsnamen>
        <urlScans>
            <urlScan id="u1">https://archief.amsterdam/inventarissen/inventaris/5075.nl.html#A15910000240.JPG
            </urlScan>
            <urlScan id="u2">https://archief.amsterdam/inventarissen/inventaris/5075.nl.html#A15910000241.JPG
            </urlScan>
            <urlScan id="u3">https://archief.amsterdam/inventarissen/inventaris/5075.nl.html#A15910000242.JPG
            </urlScan>
        </urlScans>
    </set>
</sets>
```

Figure 4: Probate inventory

5 Conclusion

This paper reports on the challenge of connecting and disambiguating person names extracted from the SAA registries during the Dutch Golden Age. Our preliminary conclusion is that,

even though the LL method is under development, *it satisfactorily meets most of our requirements for the disambiguation* task and even *supports data quality improvement*. Although most Amsterdamers are not yet disambiguated (since the validation step still needs to be taken), we reached the goal of *identifying the requirements and the difficulties to be addressed* in order to get as close as possible to disambiguate Amsterdamers. To support entity recognition of biographical and other data, and to align often fuzzy digital collections with repetitions, misspellings and spelling variations, new experiments using more data and improved versions of the LL method are relevant not only for the Golden Agents and All Amsterdam Acts projects in particular, but also for the use of linked data in the digital humanities in general.

1014 words(- 14)

References

Idrissou, A. K., Hoekstra, R., van Harmelen, F., Khalili, A., and van den Besselaar, P. (2017). Is my:sameAs the same as your:sameAs? In *Proceedings of the Knowledge Capture Conference on - K-CAP 2017*, pages 1–8, New York, New York, USA. ACM Press.