From Raw Text to Meaningful Information: Named Entity Recognition, Disambiguation, and Semantic Enrichment of a Large Corpus of Historical Police Records (Antwerp, 1876–1945)

Lith Lefranc[1]

[1] Center for Urban History, University of Antwerp, Belgium

Abstract

This paper presents a pipeline for transforming noisy, machine-readable historical records into structured, meaningful information. Unlike prior methods that often rely on contemporary natural language processing tools, this framework adapts language models and ontologies to the historical and regional specificity of the data. Using 271 incident books from Antwerp's local police (1876-1945), we developed an integrated approach combining historical named entity recognition, disambiguation, and further semantic enrichment. We trained domain-adapted transformer models on manually annotated data to extract dates, times, locations, and demographic information about individuals involved in the incidents (names, birthplaces, birthyears, and occupations). Post-processing methods address historical spelling variations, handwritten text recognition errors, and inconsistent administrative practices through normalization and disambiguation. The pipeline enriches extracted data through geocoding of historical street names using custom gazetteers, automated name-to-gender inference, and systematic conversion of occupational descriptions to social class categories via HISCO/HISCLASS standards. The system achieves F1-scores ranging from 0.82 to 0.99 across entity types, demonstrating how computational methods can unlock noisy historical records for data-driven urban history research.

Keywords: Named Entity Recognition, Historical Entity Disambiguation, Geocoding, Name-to-Gender Inference, Toponym Resolution, Night Studies, Urban History

1 Introduction

The growing digitization of archival documents opens new opportunities for historical research. However, automatically transforming vast volumes of handwritten source images into machine-readable, structured, and meaningful data remains a significant challenge. This paper presents a pipeline for processing historical police reports from Antwerp, Belgium (1876-1945), developed within the context of a broader study on social inequality in nocturnal urban life during the late nineteenth and early twentieth centuries [22].[1] This research project examines how urban modernization, such as the introduction of street lighting and intensified police surveillance, affected different social groups' access to and use of the city's nighttime spaces. Particular attention is given to four intersecting social categories: gender, class, origin, and age. These categories shape both the historical analysis and the design of the pipeline presented in this paper. To address these

Lith Lefranc. "From Raw Text to Meaningful Information: Named Entity Recognition, Disambiguation, and Semantic Enrichment of a Large Corpus of Historical Police Records (Antwerp, 1876–1945)." In: *Computational Humanities Research 2025*, ed. by Taylor Arnold, Margherita Fantoli, and Ruben Ros. Vol. 3. Anthology of Computers and the Humanities. 2025, 938–952. https://doi.org/10.63744/Aym1S8P80hvy.

[1] PhD project by Lith Lefranc, under the supervision of Ilja Van Damme and Mike Kestemont. This project is funded by the University of Antwerp (GOA FFB200403).

questions at scale, the research moves beyond traditional case studies and sampling approaches common in urban history ”from below” (e.g. [10]). Instead, it draws on a large corpus of police records, which offer systematic, street-level documentation of urban encounters over a seventy-year period.

These police records present multiple challenges typical of historical sources: inconsistent handwriting styles, evolving administrative practices, historical spelling variations, and the absence of standardized data formats. Moreover, the historical nature of the content introduces specific complexities such as obsolete place names, historical occupational titles, and changing naming conventions, which are typically overlooked by contemporary Natural Language Processing tools. The handwritten text recognition (HTR) of these records was completed in an earlier phase of the project and is documented in a separate publication [23]. Building on this automatically transcribed dataset, the present pipeline integrates named entity recognition (NER) with extensive post-processing to disambiguate and semantically enrich information from the police records into standardized and historically meaningful variables. The novelty of this approach lies in adapting NLP tools designed for contemporary language use to historical language and orthography within an end-to-end framework that combines language modeling, local gazetteers, and socio-historical classification systems to turn noisy historical text into structured evidence for social and urban history.

2 The Antwerp incident books (1876-1945)

The raw data consists of 271 so-called incident books (Dutch: “*gebeurtenisboeken*”, French: “*régistre d’événements*”) maintained by the Antwerp local police between 1876 and 1945. These, originally handwritten, records offer rich, structured descriptions of everyday urban events, including minor offenses, disturbances, and citizen-police interactions. They document not only the nature of incidents but also detailed demographic information of the individuals involved, such as name, occupations, place and date of birth, and residential address (example on Figure 1 and Table 1). Originally compiled to assist deputy commissioners in deciding whether formal charges (“*proces-verbaal*”) should be filed, these records provide a unique window into the social dynamics of a rapidly expanding city. They capture a wide array of urban experiences, from violations and domestic disputes to lost children, stray animals, and individuals found injured in public spaces, offering granular insight into the evolving relationship between citizens and municipal authorities. Although shaped by the surveillance priorities of the time, the incident books offer rare, systematic documentation of individuals from lower social strata - social groups that often left little written trace of their own [22].

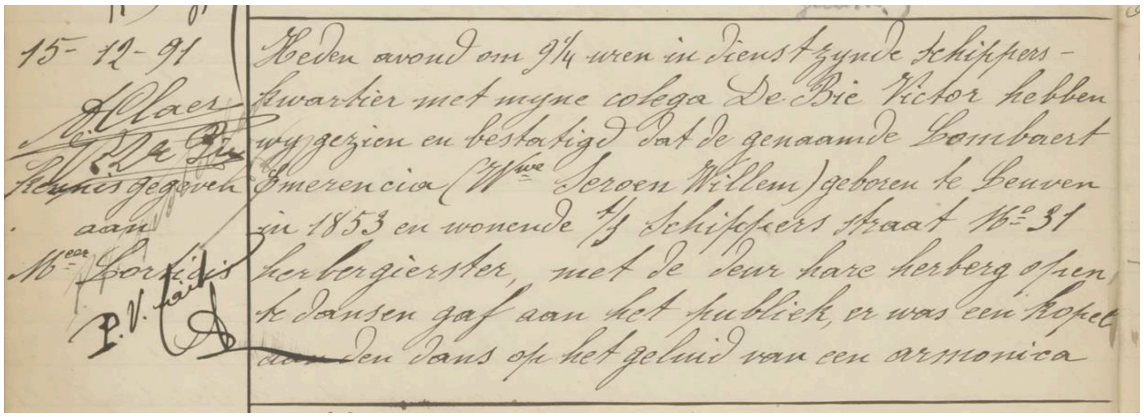

Figure 1: Example of an incident (FelixArchief, 450#58, p. 267).

The machine-readable corpus of the incident books comprises approximately 90,000 pages and more than 410,000 individual incidents, totaling over 21 million words (see Table 2). The handwritten text was automatically transcribed using ARletta [24], an open-source model for historical Dutch that combines YALTAi for object detection [7] and kraken for HTR [19]. Trained on a ground truth of 3,715 manually annotated pages, and fine-tuned with publicly available datasets, the model achieved a character accuracy rate (CAR) of 92.7%, and a relaxed CAR of 93.97% (excluding white space, punctuation, and capital letters). The automatic transcriptions provide a robust foundation for subsequent natural language processing tasks, despite fluctuations in quality due to variation in handwriting styles, page layouts, and multi-language use (approx. 94% Dutch - 6% French) [23].

Diplomatic transcription (Dutch/French)	**English Translation**
15-12-91 [signature-mark] [signature-mark] Kennis gegeven aan Mr. Lortidis. P.V. fait.	15-12-91 [signature-mark] [signature-mark] Notification given to Mr. Lortidis. Official report was made.
Heden avond om 9 1/4 uren in dienst zynde Schipperskwartier met myne colega De Bie Victor hebben wy gezien en gestatigd dat de genaamde Lombaert Emerencia (Wwe Seroen Willem) geboren te Leuven in 1853 en wonende T/S Schippersstraat N° 31 herbergierster, met de deur hare herberg open te dansen gaf aan het publiek, er was een kopel aan den dans op het geluid van een armonica	This evening around 09:30 while on duty in Schipperskwartier with my colleague De Bie Victor, we saw and confirmed that the named Lombaert Emerencia (widow of Seroen Willem), born in Leuven in 1863 and residing in this city, Schippersstraat 31, innkeeper, let the public dance in her inn with the door open. A couple was dancing to the sound of a harmonica.

Table 1: Diplomatic transcription of incident on Figure 1 (left) and non-diplomatic English translation (right) (FelixArchief, 450#58, p. 267).

Element	**Count**
Books	271
Pages (sides)	87,593
Incidents	410,829
Lines	3,804,556
Words	21,555,587
Characters	4,128,408,985

Table 2: Total count of books, pages (sides), incidents, lines, words, and characters.

3 Historical Named Entity Recognition

While the HTR pipeline successfully transformed the handwritten police records into machine-readable text, the resulting corpus remains noisy, unstructured, and difficult to analyze in its raw form. Key information like dates, locations, and personal names are embedded in free-form narrative sentences with historical spelling variations, abbreviations, and non-standard formats. To extract this information in a structured and analyzable way, we apply named entity recognition (NER), adapted to the historical and linguistic particularities of the corpus.

NER in historical documents faces unique challenges compared to contemporary applications. Language evolution can lead to ”entity drift” where entities change meaning or disappear entirely

over time [14]. Additionally, the machine-readable versions of historical texts, typically created through OCR or HTR, contain noise that complicates entity recognition. Models trained on contemporary texts often perform poorly on historical documents due to these linguistic and orthographic differences. Recent advances in transformer-based models have shown promise for historical NER tasks. While generative models are being explored, encoder-only transformers like BERT remain the standard [17].

To extract structured data from the semi-structured police reports, we developed a tailored NER approach using transformer-based language models for Dutch. We experimented with two existing models: RobBERT, trained on modern Dutch data (OSCAR-NL) [12], and GysBERT, trained on historical Dutch-language texts (Delpher and DBNL, 1500-1950) [25]. Both models were further adapted to our domain via pretraining and fine-tuning. The training data was manually annotated using the open-source platform Recogito [34]. Initially, incident book pages were randomly sampled to ensure broad coverage across neighborhoods and years. Due to fragmentation in early samples the dataset was expanded with consecutive pages from three complete incident books.[2] In total, 1,000 events were annotated, yielding 11,335 labeled entities across three levels:

- <EVENT>: Each individual incident was tagged, containing nested contextual entities:
  - <E_DATE>: 510 dates
  - <E_TIME>: 1,315 times (split into numeric times and textual phrases)
  - <E_LOC>: 931 street-level location tags
- <PERSON>: All civilians were tagged with rich internal structure (1,144 total), including:
  - <P_TITLE>: 484 titles
  - <P_FIRSTNAME>: 1,333 firstnames
  - <P_LASTNAME>: 2,340 lastnames
  - <P_PROF>: 542 occupations
  - <P_BIRTHPLACE>: 522 birthplaces
  - <P_BIRTHDATE>: 471 birthdates
  - <P_AGE>: 146 ages
  - <P_RES>: 788 places of residence
  - <P_MARITALSTATUS>: 77 marital statuses (including kinship terms like daughter or sister)
- <P_OFFICER>: To avoid confusion between civilians and police officers, each police officer was annotated separately (1,203 total), including their firstname, lastname and title.

The models were adapted in two stages. First, we applied domain-specific pretraining on the full HTR corpus of the incident books using masked language modeling. This allowed both RobBERT and GysBERT to internalize the specific syntax, spelling, and vocabulary of the source material. Second, we added a token-level NER classification head, trained on the annotated data using the BIO scheme. Both a coarse-grained model (detecting only <PERSON> and <P_OFFICER>) and a fine-grained model (with all 12 sub-entity types) were trained.

[2] FelixArchief, 731#1613, p.191-240; 450#58, p.89-138; MA#30825, p.30-79.

Model	Dataset	Precision	Recall	F1
robbert-2023-dutch-large-adapted	NER-fine	0.786	0.855	0.819
GysBERT-adapted	NER-fine	0.779	0.858	0.817
GysBERT-adapted	NER-coarse	0.708	0.805	0.754
robbert-2023-dutch-large-adapted	NER-coarse	0.646	0.779	0.706

Table 3: General evaluation of adapted NER models.

On the fine-grained task, RobBERT achieved F1: 0.819, while GysBERT scored F1: 0.817. On the coarse-grained task, GysBERT performed slightly better (F1: 0.754) than RobBERT (F1: 0.706), though RobBERT had a notably higher precision (see Table 3). Across all tasks, recall exceeded precision, indicating a tendency to over-identify entities - beneficial for recall-critical use cases but requiring careful post-processing. Per-entity scores (see Figure 3) revealed that RobBERT consistently performed better for most entity types, particularly on name-related fields crucial for disambiguation and normalization tasks. Its high precision made it preferable for downstream use, despite its modern training base. Its strong performance can likely be attributed to the relative recency (1876-1945) of the corpus and effective domain adaptation. The fine-tuned RobBERT model (robbert-2023-dutch-large-adapted) was selected for further use in this pipeline. The structured output (example see Table 2) forms the basis for the following steps: downstream normalization, disambiguation, and semantic enrichment.[3]

```
<incident number="2">
<marg number="2.1">
<E_DATE>15 - 12 - 91</E_DATE> Klaet
Dr Bu Keynis gegeven M aan ed Jortidis
</marg>
<p number="2.2">
Heden <E_TIME>avond</E_TIME> om <E_TIME>9 1 / 4 uren</E_TIME>
in dienst zynde <E_LOC>schippers kwartier</E_LOC>
met myne colega <P_OFFICER><P_LASTNAME>De Bie</P_LASTNAME>
<P_FIRSTNAME>Victor</P_FIRSTNAME></P_OFFICER>
hebben wy gezien en bestatigd dat de genaamde
<PERSON><P_LASTNAME>Lombaert</P_LASTNAME>
<P_FIRSTNAME>Imerencia</P_FIRSTNAME>
(<P_MARITALSTATUS>Wwe Seroen</P_MARITALSTATUS> Willem)
geboren te <P_BIRTHPLACE>Leuven</P_BIRTHPLACE> in
<P_BIRTHDATE>1853</P_BIRTHDATE> en wonende
<P_RES>t / s Schippers straat N 31</P_RES>
<P_PROF>herbergierster</P_PROF></PERSON>,
met de deur hare herberg open te dansen gaf aan het publiek,
er was een kopel aan den dans op het geluid van een armonica
</p>
</incident>
```

Figure 2: Example of an automatically transcribed and annotated incident with NER models robbert-coarse and robbert-fine (450#58, p. 267).

[3] Both models are available at: `https://huggingface.co/emanjavacas/robbert-2023-dutch-large-adapted-coarse.ner` and `https://huggingface.co/emanjavacas/robbert-2023-dutch-large-adapted-fine.ner`

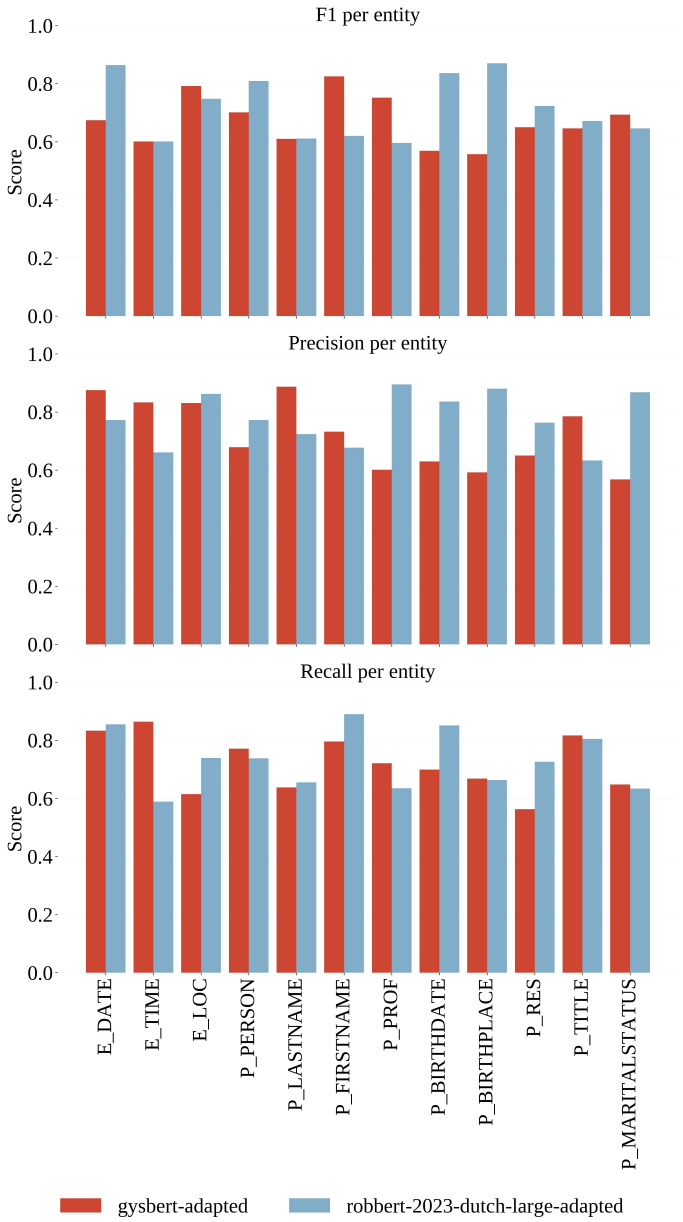

Figure 3: F1, precision, and recall per entity of gysbert-adapted and robbert-2023-dutch-large-adapted.

4 Entity Normalization, Disambiguation and Semantic Enrichment

The raw output of the NER is not yet suitable for analysis. Variability in spelling, writing style, and HTR errors, as well as a wide range of possible values for each entity type, require further normalization, disambiguation, and semantic enrichment. In the following sections, we describe the intermediate steps taken to standardize key entity types and prepare them for structured analysis: date (4.1), time (4.2), street name (4.3), gender (4.4), occupation (4.5), birthplace (4.6), and age (4.7). We used an annotated sample of 200 randomly selected incidents throughout the entire corpus as ground truth to assess the accuracy of our post-processing tools.

While standardization is necessary for computational analysis, it also involves categorization - a process that reduces linguistic richness to predefined, limited classes. These classifications are not neutral but shaped by historical and contemporary power structures that determine what and

who is considered relevant [30]. This is especially true for sources produced by authorities, such as police officers. We aim to be transparent about the methodological choices behind our categories and transformations, recognizing their analytical utility as well as their limitations.

4.1 Date Normalization

Dates extracted via NER (<E_DATE>) from the incident books displayed considerable variation in format due to inconsistent police notation, historical spelling, and HTR noise. Examples range from numeric forms (e.g. "19-8-04", "30.8.04") to written variants like "28 oogst 04" or "9den dezen". To enable temporal analyses like seasonal patterns, daylight estimation, and age calculation, all recognized dates were normalized to a uniform dd/mm/yyyy-format. This process involved regular expression patterns to detect a broad range of date formats, parsing each into discrete components (day, month, year). While days and years were typically numeric and straightforward to parse, months occurred in diverse formats: numeric, full names (e.g. "januari"), abbreviations, or even archaic terms (e.g. "oogst", "Xber"). To map these to standard month numbers, we combined dictionary-based matching with fuzzy matching (using Levenshtein distance), allowing tolerance for minor HTR or spelling errors. Parsed dates were then validated and corrected where possible. This included logical checks (e.g. rejecting "31 April" or "29 February" in non-leap years) and boundary checks using the metadata of each incident book to flag implausible years. Missing or partially parsed values were completed using context: either the last valid date mentioned or the book's metadata. Because incidents are often grouped under a single heading date in the books, undated events inherit the last recognized date. This context-aware strategy ensured near-complete date coverage.

Evaluation on the manually annotated test set showed high performance: precision reached 0.98, recall 1.00, and F1-score 0.99. Errors occurred only in rare cases due to HTR distortion of day or month components.

4.2 Time Normalization and Categorization

Time expressions (E_TIME) in the incident books, though more systematic than dates, still displayed considerable variation. Officers consistently used the 12-hour system, but annotated time in both numeric and fractional formats (e.g., "945", "9 3/4"), often accompanied by daypart indicators (e.g., "'s morgens" (EN : in the morning), "'s avonds" (EN : in the evening)). To make these expressions usable for analysis, all recognized times were normalized to a standard 24-hour format (hh:mm). The raw time data was first cleaned by removing superfluous spaces and harmonizing fractional notations. Composite expressions like "9 3/4 avond" (EN : 9 3/4 in the evening) were split into numeric and textual segments. Dayparts were matched against multilingual dictionaries (Dutch, French) and detected using tolerant regular expressions that accounted for HTR errors and historical variants (e.g., "'s avonds"). This was crucial for correctly disambiguating morning vs. evening times in the 12-hour system. Without a recognizable daypart, the time was marked as unknown.

Numeric segments were parsed as full hours, fractions (e.g., "1 3/4"), or combinations with minutes (e.g., "5 ure 20 minuten"). Fractions were converted into minutes and all values validated to ensure internal consistency (e.g., 23:59 maximum). Each parsed time was then checked against its associated daypart for logical alignment. If a mismatch was found (e.g., "2:30" with "voormiddag"), or if the numeric segment was unreliable, a default value for the detected daypart was assigned (e.g., "voormiddag" = 09:00). When only a daypart was available, the default time was used directly. Evaluation on the manually labeled test set yielded a precision of 0.99, recall of 0.81, and an F1-score of 0.89. Most errors stemmed from missing or unrecognized dayparts, highlighting the model's cautious approach: only sufficiently certain values were retained.

To enable higher-level temporal analysis, normalized times were categorized in two ways:

- **Astronomical time** Using PyEphem, a python library for astronomical calculations [31], we calculated the sun position for each event in Antwerp (lat 51.22, lon 4.40), assigning it to one of four zones: civil dawn, daylight, civil dusk, and night [33]. We accounted for historical timekeeping practices: pre-1892 events used local solar time (UTC+00:17:30), while later ones used national standards (Greenwich Mean Time, then CET during WWII) [2]. This approach captures seasonal variability: e.g., civil twilight in summer lasts 50 minutes vs. 35 minutes in spring/fall.
- **Socio-cultural time** Building on historical labor and mobility studies [11; 28; 29], we divided the day into fixed blocks reflecting common urban routines (Table 4). For instance, 18:00-21:00 was marked as "evening" and 0:00-3:00 as "night". These periods reflect working-class schedules shaped by industrialization, gas lighting, and labor reforms (e.g., 8-hour workday introduced in 1924). The categorization is wide enough to accommodate social diversity, including women, children, and the elderly, whose street activity followed different rhythms.

Time Period	Category
6:00-9:00	Morning
9:00-12:00	Late morning
12:00-15:00	Afternoon
15:00-18:00	Late afternoon
18:00-21:00	Evening
21:00-0:00	Late evening
0:00-3:00	Night
3:00-6:00	Late night/early morning

Table 4: Socio-cultural time categories.

Combining astronomical and socio-historical perspectives on time enables us to study whether natural light cycles or cultural routines better explain when and how people appeared in public space during this period.

4.3 Geocoding Street Names

To understand the spatial impact of urban modernization on Antwerp's nocturnal street life, the ability to localize incidents is essential. This requires geocoding: converting textual location descriptions into geographic coordinates [15]. However, geocoding historical street names presents specific challenges: street names often change over time, and entire streets may be relocated or disappear due to urban restructuring. Popular modern tools like Google Maps or OpenStreetMap's Nominatim cannot handle these historical variations.

For historical Antwerp, the GIStorical Antwerp platform offers a solution [18]. It provides georeferenced street data for seven historical snapshots of the city (1584-1984). For our project, the 1898 layer was used as a base [35]. Yet, because the incident books cover a wider period, many recognized street names did not appear in this dataset. To address this, the reference layer was expanded using Robert Vandeweghe's *Geschiedenis van de Antwerpse straatnamen* [36]. The combined dataset grew from 1,184 to 2,351 distinct georeferenced entries. Frequently mentioned non-street locations (e.g. train stations or hospital) were also added manually.

The next step was to match <E_LOC> entities resulting from the NER to this enriched list. First, the recognized street names were normalized through preprocessing: converted to lowercase, stripped of articles and punctuation, corrected for abbreviations (e.g., ”str” → ”straat”), and stripped of spacing artifacts introduced by HTR. The reference list was cleaned using the same procedure to enable consistent comparison. Next, matching was done using fuzzy string matching (Levenshtein distance). Match scores were standardized to a 0-100 scale. Based on heuristic testing, matches scoring above 70 were accepted; others were labeled ”unknown”. Evaluation on the annotated test set showed strong results: precision 0.98, recall 0.91, and an F1-score of 0.94. Errors were mainly caused by missing entries in the reference list or HTR/NER distortions.

4.4 Name-to-Gender Inference

Because the incident books do not explicitly record the gender of individuals, gender must be inferred from indirect indicators, most notably first names. Over the past years, first-name-based gender inference has become a standard method in computational research [32], yet it is inherently limited, especially in historical contexts. Most tools rely on binary (male/female) classifications and cannot account for non-binary or culturally fluid gender identities [5; 6; 8]. While these categories are certainly present in society, they were not institutionally recognized or recorded in late 19th- and early 20th-century police records. Moreover, these inference models often assume a stable association between names and gender across time and place, which is rarely the case. Naming practices are historically and regionally contingent: names such as ”Leslie” or ”Maria” may shift in gendered connotation across decades, cultures, or social classes [3]. In addition, many inference tools are trained on unbalanced datasets that underrepresent women - a phenomenon described as the “gender data gap” [26]. This systematic bias, rooted in the historical treatment of male experiences as normative, results in tools that often perform better for male-associated names than for female ones. Compounding this issue, most available tools are trained on contemporary or Anglophone datasets, which makes them prone to error when applied to Dutch- and French-language historical data. Given these challenges, gender in this project is not treated as an objective or self-reported attribute but as a historically imposed and analytically inferred category. The resulting label `gender_inferred` is explicitly marked as such to emphasize its constructed nature and methodological limitations.

Initially, we applied the open-source tool gender-guesser [27], which assigns binary and probabilistic gender labels to names based on a curated list of 45,000 entries, grouped by country. Despite its relatively balanced dataset and international scope, gender-guesser is based on data from 2007 and struggles with historical or region-specific names. On our sample test set, it achieved an F1-score of 0.87, with very high precision (0.99) but lower recall (0.77), due to its frequent labeling of historical names or names with HTR errors as “unknown” (e.g., “Seraphine”,“Daoid”).

To improve recall and better match the historical and regional context, we developed a custom tool: historical-Antwerp-gender-predictor. This firstname-gender dictionary is based on a large-scale dataset from Antwerp’s civil death registers (1820-1946), which contains firstnames and registered gender of 484,596 individuals (259,004 men; 225,592 women).[4] The dataset was used to compute gender frequencies for 6,746 unique first names. Each name was assigned a gender if one occurred in more than 80% of cases; otherwise, it was marked as “uncertain”. In applying this tool to the incident book data, we used only the first listed given name, as subsequent names often vary in gender or reflect familial inheritance (e.g., “Maria” in male name clusters). Names were normalized (e.g., case folding, removal of initials/special characters), and matched to the dictionary using fuzzy matching (Levenshtein distance, max. difference of two characters)

[4] This dataset was developed in the context of the research project S.O.S. Antwerp, which investigates social inequality in cause-specific mortality in Antwerp between 1820 and 1946 [13].

to accommodate minor HTR errors (e.g., “Mria” → “Maria”). If no match was found, the gender was set to “unknown”. Where available, titles or marital status (e.g., “Madame”, “dochter” (EN: daughter), “weduwe” (EN: widow)) were prioritized over name inference to reduce false positives, especially in cases where women were listed under a husband’s or father’s name (e.g., “Mme Henri Louis Backer”).

The historical-Antwerp-gender-predictor outperformed gender-guesser with an F1-score of 0.98, owing to both precision (0.98) and a markedly higher recall (0.98). Most remaining errors stemmed from severe HTR distortion. Due to its superior performance and contextual accuracy, historical-antwerp-gender-predictor was selected as the final gender inference method.

4.5 Occupation to Social Class

As with gender, the incident books do not explicitly record socio-economic status (SES), but occupations (<P_PROF>) offer an indirect indication. Using occupation as a proxy for SES presents however some challenges. It is a less precise measure than income or property, it underrepresents women (due to lower labor participation and occupational variety), and it excludes children and the elderly. In addition, the meaning and social prestige of occupations change over time, and many historical occupations no longer exist today. Nevertheless, occupation is one of the few systematically recorded attributes in the incident books that offer insight into social position, making it a valuable, yet imperfect, indicator of social class.

To enable diachronic and transnational comparison, the Historical International Classification of Occupations (HISCO) system was used [21]. HISCO groups historical occupations into a hierarchical structure of 1,675 detailed codes, modeled on ISCO-68, and based on the tasks involved in each job. To measure social stratification, HISCO was linked to HISCLASS [20], which classifies HISCO occupations into social classes based on skill level, manual or non-manual labor, supervision, and economic sector [20].

We first applied OccCANINE to assign HISCO codes to the extracted occupations, which is a multilingual classifier trained on 14 million labeled job descriptions [9]. Applied to both Dutch and French entities, it achieved an F1-score of 0.80 (precision 0.78; recall 0.82). Main issues stemmed from HTR noise (e.g., “touwersgast” and from occupations specific to 19th-century Antwerp (e.g., “lichtmeisje” (EN: sex worker), “dokwerker” (EN: dockworker)) or non-economic statuses like “scholier” (EN: student). To improve the results, we built a custom HISCO-HISCLASS dictionary, combining official HISCO codes with local, historical, and non-standard occupations. Using fuzzy matching (Levenshtein distance ≤ 0.3), each <P_PROF> entity was linked to its closest match in the dictionary. Preprocessing included lowercasing, removing whitespace, and standardizing accents. This custom method maintained the same recall (0.82) as OccCANINE but substantially improved precision (0.92), resulting in a higher F1-score of 0.87. It better handled noisy inputs (“touwersgast” → “brouwersgast” (EN: brewer’s servant) and successfully matched local terms like “lichtmeisje”. However, it lacked contextual awareness; for instance, “dekwerker” was misclassified as “dakwerker” (EN: “roofer”) instead of “dokwerker”, distorting both HISCO and HISCLASS results. Despite these limitations, the custom approach proved more accurate and better adapted to this noisy dataset.

4.6 Birthplace Resolution

Toponym resolution, the linking of place names to specific geographic coordinates, consists of two key steps: toponym recognition and disambiguation. While modern systems use gazetteers like GeoNames or Wikipedia-based knowledge graphs, these are ill-suited for historical data, which contain defunct or renamed places and greater spelling variation due to language change, transcription errors, and inconsistent usage. To disambiguate birthplaces (P_BIRTHPLACE) in the in-

cident books, we used the World Historical Gazetteer (WHG), which links historical place names to coordinates and time periods [16]. As WHG lacks a public API, we requested a TSV export for European toponyms. This dataset required cleaning due to inconsistencies and duplicates, because many entries stem from user uploads with varying coordinate precision. The cleaning steps included: normalizing all toponyms to lowercase, removing entries with only numeric values, rounding coordinates to two decimals (1 km precision), removing duplicates based on names with near-identical coordinates, and manually adding expressions like "T/S" (abbreviation for "ter stede" (EN: in this city)) and "alhier" (EN: here) as references to Antwerp.

The birthplaces extracted from the incident books were preprocessed (lowercasing, removal of punctuation, digits, and extra spacing) and then matched using fuzzy string matching (Levenshtein distance) and Haversine distance from Antwerp. For ambiguous fuzzy matches (Levenshtein ratio between 0.70 and 0.85), candidates were re-ranked by geographic proximity. For instance, "Berghem" would match to Berchem (Belgium) over Berghem (Netherlands) due to both lexical similarity and spatial nearness. If multiple candidates had identical scores, the top result was selected for consistency. If no suitable match was found, the birthplace field was left empty.

On the test set, the toponym resolution method achieved an F1-score of 0.82, with precision at 0.76 and recall at 0.89. The majority of errors stemmed from HTR noise, while others were due to incorrect geographic ranking (e.g., "Gronbeck" mislinked to "Grotenbroeck" instead of "Groesbeek"). Although police agents sometimes noted clarifying info (e.g., country names in brackets), this was too inconsistent to systematically integrate into the algorithm. Despite these limitations, the approach provided sufficiently reliable spatial grounding for birthplace data across the corpus.

4.7 Age Calculation and Categorization

The incident books do not systematically record individuals' ages, but age-related information is often indirectly available via birthdates (<P_BIRTHDATE>) or explicit age mentions (<P_AGE>). Because age and birthdate expressions vary widely, ranging from "37 jaar" (EN: 37 years) to "13 october 1836", both variables were normalized into a single numeric age value (in years). Birthdates were converted to a standardized dd/mm/yyyy-format using regular expressions, from which the birth year was extracted. For two-digit years, the correct century was determined by calculating the difference with the year of the event. Age strings were cleaned as follows: references to "months" were recoded as 0; vague mentions like "kind" (EN: child) were coded as 5; and in composite numbers, only the first numeric value was retained. When only a birth year was available, age was calculated by subtracting the birth year from the event year. A series of validity checks was then applied: valid birth years fell between 1750 and 1950; valid ages ranged from 0 to 100. Additional plausibility rules flagged inconsistencies, e.g., individuals under 5 or over 80 with an occupational code were marked as invalid, assuming it unlikely they were economically active. The age normalization script achieved an F1-score of 0.94 (precision: 0.98, recall: 0.92). Remaining errors were primarily due to upstream NER or HTR mistakes.

Crucially, numerical age alone does not suffice for historical analysis. As Philippe Ariès emphasized, age is both biologically and culturally constructed, shaped by shifting norms around schooling, marriage, work, and retirement [1]. In this period, age was not experienced or institutionalized uniformly: class and gender shaped access to education, entry into the labor market, marriage timing, and overall life expectancy. For example, men from upper classes typically spent more years in school than working-class men or women, and the average age at marriage for men was generally higher than for women. These intersecting factors influenced not only life trajectories but also how age was socially interpreted [4]. Since lower social classes are overrepresented in the incident books, the age categories were defined with particular attention to their life course patterns. To reflect these historical realities, we divided age into five analytically meaningful life

stages (see Table 5), based on typical schooling age, working life, marriage age, and life expectancy during the period under study.

Age	Age category	Context
0-4	Young child	before school-age
5-14	Child	school-age
15-24	Young adult	professionally active, but not yet married
25-64	Adult	professionally active and/or married
65+	Elderly	average life expectancy

Table 5: Age categories based on historical context.

5 Conclusion

This paper demonstrates how historical police records, which are often considered too inconsistent and noisy for computational analysis, can be systematically transformed into structured, semantically enriched information. By developing a tailored pipeline for the Antwerp police incident books, we combined handwritten text recognition, historical named entity recognition, and domain-specific post-processing to extract historically and spatially contextualized information about individuals encountered by the police in Antwerp's streets. Our results show that transformer-based language models, when properly adapted to historical data, achieve high performance despite orthographic variation and HTR noise. The fine-tuned RobBERT model achieved a macro F1-score of 0.819 for entity recognition, while post-processing normalization and disambiguation steps consistently performed above 0.82, with several reaching near-perfect scores (0.94-0.99). Custom tools like the historical-Antwerp-gender-predictor and localized HISCO-HISCLASS mappings proved essential for handling historical and regional specificity, consistently outperforming generic alternatives.

The resulting structured dataset covers over 400,000 incidents with detailed temporal, spatial, and socio-demographic information, and opens new avenues for digital urban history. For the broader doctoral research project on social inequality in nocturnal Antwerp, this enriched dataset enables systematic analysis of who had access to nighttime urban spaces and how urban modernization affected the social geography of the city after dark. The ability to analyze hundreds of thousands of street-level encounters exemplifies the potential of scalable reading approaches in historical research. The pipeline's modular design makes it adaptable to other historical sources with similar characteristics, providing a template for projects seeking to bridge archival abundance and analytical accessibility. However, we emphasize the epistemological limits of this computational approach. Categories like binary gender classifications and hierarchical social classes, while necessary for analysis, risk flattening the complexity of historical lived experience. Police records inherently overrepresent interactions with marginalized groups and carry institutional biases that our methods inevitably reproduce. Ultimately, this work illustrates how computational methods combined with historical sensitivity can unlock large-scale archival sources while remaining attentive to their limitations. The true test lies not in technical performance metrics, but in the capacity to generate new historical insights grounded in careful interpretation of the contexts that shaped both the original records and our analytical choices.

Acknowledgements

I would like to thank Ilja Van Damme and Mike Kestemont for supervising this PhD project, Enrique Manjavacas and Pieter Fivez (TEXTUA) for their help with the NER model training, and

Folgert Karsdorp, Melvin Wevers, and the reviewers of this paper for their thoughtful comments and suggestions.

References

[1] Ariès, Philippe. *Histoire des populations françaises*. Paris: Editions Self, 1948.

[2] Belgium, Royal Observatory of. "Time and daylight saving time". 2025. URL: https://robinfo.oma.be/en/astro-info/time/.

[3] Blevins, Cameron and Mullen, Lincoln. "Jane, John … Leslie? A Historical Method for Algorithmic Gender Prediction". In: *Digital Humanities Quarterly 9* (2015). URL: https://www.digitalhumanities.org/dhq/vol/9/3/000223/000223.html.

[4] Bourdelais, Patrice and Gourdon, Vincent. "Demographic categories revisited. Age categories and the age of categories". In: *Human Clocks: The Bio-cultural Meanings of Age*, ed. by Claudine Sauvain-Dugerdil, Henri Léridon, and Christopher G. Nicholas Mascie-Taylor. Bern: Peter Lang, 2006, pp. 245–269.

[5] Butler, Judith P. *Bodies That Matter: On the Discursive Limits of Sex*. New York: Routledge, 1993.

[6] Butler, Judith P. *Gender Trouble: Feminism and the Subversion of Identity*. New York: Routledge, 1990.

[7] Clérice, Thibault. "You Actually Look Twice At it (YALTAi): Using an object detection approach instead of region segmentation within the Kraken engine". In: *Journal of Data Mining & Digital Humanities* (2023). DOI: 10.46298/jdmdh.9806.

[8] D'Ignazio, Catherine and Klein, Lauren F. *Data Feminism*. Cambridge, Massachusetts: The MIT Press, 2020.

[9] Dahl, Christian Møller, Johansen, Torben, and Vedel, Christian. "Breaking the HISCO Barrier: Automatic Occupational Standardization with OccCANINE". 2024. DOI: 10.48550/arXiv.2402.13604. eprint: arXiv.

[10] De Koster, Margo. "Negotiating Controls, Perils, and Pleasures in the Urban Night: Working-Class Youth in Early-Twentieth-Century Antwerp". In: *Criminological Encounters 3* (2020), pp. 32–49. DOI: 10.26395/CE20030104.

[11] De Vos, Patrick. "De historische evolutie van de arbeidsduur in België". In: *Brood & Rozen 2*, no. 3 (1997), pp. 7–37. DOI: 10.21825/br.v2i3.2680.

[12] Delobelle, Pieter, Winters, Thomas, and Berendt, Bettina. "RobBERT: a Dutch RoBERTa-based Language Model". 2020. DOI: 10.48550/arXiv.2001.06286. eprint: arXiv.

[13] Devos, Isabelle and Greefs, Hilde. "Sociale Ongelijkheid in Sterfte Antwerpen (1820-1946)". 2021. URL: https://sosantwerpen.be/.

[14] Ehrmann, Maud, Hamdi, Ahmed, Pontes, Elvys Linhares, Romanello, Matteo, and Doucet, Antoine. "Named Entity Recognition and Classification in Historical Documents: A Survey". In: *ACM Computing Surveys 56*, no. 2 (2023), pp. 1–47. DOI: 10.1145/3604931.

[15] Goldberg, Daniel W., Wilson, John P., and Knoblock, Craig A. "From Text to Geographic Coordinates: The Current State of Geocoding". In: *URISA Journal 19*, no. 1 (2007), pp. 33–46. DOI: 10.1002/9781118786352.wbieg1051.

[16] Grossner, Karl, Grunewald, Susan, and Mostern, Ruth. "Bringing Places from the Distant Past to the Present: A Report on the World Historical Gazetteer". In: *International Journal on Digital Libraries 24*, no. 3 (2023), pp. 159–162. DOI: 10.1007/s00799-022-00341-2.

[17] Hiltmann, Torsten et al. “NER4all or Context is All You Need: Using LLMs for low-effort, high-performance NER on historical texts. A humanities informed approach”. 2025. DOI: 10.48550/arXiv.2502.04351. eprint: arXiv.

[18] Jongepier, Iason and Janssens, Ellen. “GIStorical Antwerp: historisch GIS als laboratorium voor de stadsgeschiedenis”. In: *Stadsgeschiedenis 10* (2015), pp. 49–62.

[19] Kiessling, Benjamin, Tissot, Robin, Stokes, Peter, and Ezra, Daniel Stökl Ben. “eScriptorium: An Open Source Platform for Historical Document Analysis”. In: *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*. Vol. 2. 2019. DOI: 10.1109/ICDARW.2019.10032.

[20] Leeuwen, Marco H. D. van, Maas, Ineke, and Miles, Andrew. “Creating a Historical International Standard Classification of Occupations: An Exercise in Multinational, Interdisciplinary Cooperation”. In: *Historical Methods 37* (2004), pp. 186–197. DOI: 10.3200/HMTS.37.4.186-197.

[21] Leeuwen, Marco H. D. van, Maas, Ineke, and Miles, Andrew. *HISCO: Historical International Standard Classification of Occupations*. Leuven: Universitaire Pers Leuven, 2002.

[22] Lefranc, Lith. “Nieuw licht op de stedelijke nacht. Digitaal speuren naar ’onzichtbare flaneurs’ in de gebeurtenisboeken van Antwerpen (1876-1939)”. In: *Stadsgeschiedenis 20*, no. 1 (2025), pp. 65–82.

[23] Lefranc, Lith, Kestemont, Mike, and Van Damme, Ilja. “Antwerp Street Stories. ±90,000 machine-readable pages of handwritten local police reports (1876-1945)”. In: *Research Data Journal for the Humanities and Social Sciences 1* (2025). in publication.

[24] Lefranc, Lith, Van Damme, Ilja, Clérice, Thibault, and Kestemont, Mike. “ARletta. Open-Source Handwritten Text Recognition Models for Historic Dutch”. In: *Journal of Open Humanities Data 10* (2024), p. 43. DOI: 10.5334/johd.225.

[25] Manjavacas Arevalo, Enrique and Fonteyn, Lauren. “Non-Parametric Word Sense Disambiguation for Historical Languages”. In: *Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities (NLP4DH 2022)*. Association for Computational Linguistics, 2022, pp. 123–134. DOI: 10.18653/v1/2022.nlp4dh-1.16.

[26] Perez, Carolina Criado. *Invisible Women: Exposing Data Bias in a World Designed for Men*. New York: Vintage Books, 2020.

[27] Perez, Israel Saeta. “gender-guesser: Get the gender from first name”. 2016. URL: https://github.com/lead-ratings/gender-guesser.

[28] Pierik, Bob. *Urban life on the move. Gender and mobility in early modern Amsterdam*. PhD thesis. Amsterdam: University of Amsterdam, 2022.

[29] Pooley, Colin. “On the street in nineteenth-century London”. In: *Urban History 48*, no. 2 (2021), pp. 211–226. DOI: 10.1017/S096392681900097X.

[30] Posner, Miriam. “What’s Next: The Radical, Unrealized Potential of Digital Humanities”. Blog post. 2015. URL: https://miriamposner.com/blog/whats-next-the-radical-unrealized-potential-of-digital-humanities/.

[31] Rhodes, Brandon. “PyEphem”. 2021. URL: https://rhodesmill.org/pyephem/.

[32] Santamaría, Lucía and Mihaljević, Helena. “Comparison and benchmark of name-to-gender inference services”. In: *PeerJ Computer Science 4* (2018), e156. DOI: 10.7717/peerj-cs.156.

[33] Schilling, Govert. *The Astronomy Handbook: The Ultimate Guide to Observing and Understanding Stars, Planets, Galaxies, and the Universe*. New York: Black Dog & Leventhal, 2024.

[34] Simon, Rainer, Barker, Elton, Isaksen, Leif, and De Soto CaÑamares, Pau. “Linked Data Annotation Without the Pointy Brackets: Introducing Recogito 2”. In: *Journal of Map & Geography Libraries 13*, no. 1 (2017), pp. 111–132. DOI: 10.1080/15420353.2017.1307303.

[35] Van Damme, Ilja, Janssens, Ric, Jongepier, Iason, Klaarenbeek, Reinout, Kooten, Rogier van, and Hermenault, Léa. “GIStorical Antwerp: Micro-level layers: layer of 1898”. 2020. URL: https://www.uantwerpen.be/en/projects/gistorical-antwerp/about-the-project/microlevellayers/.

[36] Vandeweghe, Robert. *Adresboek van de stad en de provincie Antwerpen*. Antwerpen: Mercurius, 1997.