All Present, None Accounted For: Named Entity Extraction as Diagnostic for Long-Tail Knowledge in Historical Archives

Jude Majumdar[1], Delfi Sol Martinez Pandiani[1]

[1]University of Amsterdam, Faculty of Humanities, Amsterdam, The Netherlands

Abstract

Historical notarial records capture everyday legal and economic transactions involving a wide range of social actors, offering rich evidence about ordinary lives and relationships. Yet access is typically mediated by provenance-based finding aids that organize holdings by record creator, document type, and date range, rather than by the persons documented in the records. This model clashes with emerging archival standards that call for entity- and relation-centered description. We use the Amsterdam City Archives' fonds 5075—over 20,000 volumes of notarial deeds spanning 1578–1915—as a case study of how far state-of-the-art NLP can recover long-tail person entities and relations from early modern European notarial archives. Using Handwritten Text Recognition (HTR) transcriptions of 2,282 volumes, we apply Named Entity Recognition (NER) to extract 1.5 million unique person-name strings and analyze the resulting population as structurally long-tail across social, linguistic, and temporal dimensions. Socially, it consists overwhelmingly of ordinary actors—including artisans, widows, and creditors—whose lives often leave few traces beyond notarial deeds (82% of extracted names appear in only one inventory). Linguistically, early modern Dutch patronymic naming yields short, structurally thin name strings with higher NER recall but zero authority file matches, whereas structurally richer multi-part names have lower NER recall but higher authority matching (8.2%). Temporally, the seventeenth and eighteenth centuries dominating this corpus lie largely outside the focus of most digital authority files and knowledge bases. We thus find that provenance-based descriptions, name indexes, and authority files each independently fail to represent most of the people recorded in the deeds. We address this descriptive invisibility by implementing typed relation extraction over formulaic notarial language, recovering 44,249 familial relations at 100% precision on a 794-sample validation. The resulting structured output forms a preliminary layer for long-tail knowledge graph construction aligned with emerging archival standards. Our alignment framework and extraction pipeline are transferable to other hierarchically organized archives, offering archivists, digital humanists, and NLP practitioners a method for bringing long-tail notarial populations into view.

Keywords

long-tail knowledge, Named Entity Recognition, archival description, notarial archives, knowledge graph construction, historical text

1. Introduction

In early modern European cities, notaries recorded everyday legal and economic transactions such as wills, property sales, marriage contracts, and debt agreements in documents called deeds. These notarial deeds, preserved today in municipal archives, contain dense, structured information about past people and their relationships: they name artisans, widows, merchants, servants, and creditors alongside elites, often in highly regularized legal language. Much of this information concerns ordinary actors who sit in the long tail: they are poorly documented outside these records, hard to search, and rarely represented in authority files or knowledge graphs. Making such long-tail populations visible is increasingly recognized as a core archival task. Recent archival quality initiatives such as the Records in Contexts (RiC) standard, for example, call for archival description organized around entities and their relationships rather than record creators [1], yet large historical collections still lack entity-based representations at scale.

The 25th International Conference on Knowledge Engineering and Knowledge Management (EKAW 2026) - Satellite Events (Workshops, Tutorials, Demos, Posters), Turin, Italy

jude.majumdar@student.uva.nl (J. Majumdar); d.s.martinezpandiani@uva.nl (D. S. M. Pandiani)

https://orcid.org/0000-0003-2392-6300 (D. S. M. Pandiani)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

published 2026-11-26

CEUR-WS.org/Vol-4340/paper18.pdf

To explore what entity-based access to long-tail populations might look like in practice, we take one of the largest surviving notarial collections in Europe—the Amsterdam City Archives' (Stadsarchief Amsterdam, hereafter SAA) fonds 5075, comprising over 20,000 volumes covering 1578 to 1915—as a test case for recovering person entities and relations from archival records and their access infrastructure. Access to fonds 5075 is mediated through a hierarchical finding aid structured by *provenance*, the principle that records are grouped by who produced them rather than by what they contain [2], privileging record creators over the long-tail of ordinary actors documented in the deeds. The finding aid organizes the fonds by notary into 2,069 entries, typically consisting of a document type plus date range (e.g., "*Minuutacten*,[1] 1730–1774") without naming any of the people recorded in those deeds. Additional access layers include HTR transcriptions produced using Transkribus [3], crowdsourced person name annotations from VeleHanden [4] (1.6 million tags across nearly 300,000 deeds), and external authority files including the SAA's Person Reconstructions and ECARTICO, a prosopographical database covering the cultural industries of the Low Countries ca. 1475–1725 [5].

Researchers approaching fonds 5075 without a specific name, whether to investigate the social composition of a notarial practice, trace creditor networks across a period, or identify populations active in a particular type of transaction, have no entry point through any of these access layers. Drawing on Bowker and Star [6], we propose *descriptive invisibility of archival populations* as a form of long-tail access failure: a condition in which a population documented in archival records is unreachable through the access infrastructure that mediates research into those records, because each layer's organization independently precludes that population's representation. Unlike *archival silences*, gaps produced by decisions about what to preserve or collect [7, 8], descriptive invisibility concerns materials already held and fully preserved; it is produced by how those materials are described and organized for access.

We approach the person entities documented in fonds 5075 as constituting long-tail knowledge across three dimensions. Socially, the population consists overwhelmingly of ordinary actors (artisans, widows, creditors) who generated few archival traces beyond notarial deeds and are absent from biographical databases. Linguistically, early modern Dutch patronymic naming produces name strings that are expected to be too structurally thin to support disambiguation or authority file matching. Temporally, the 17th and 18th centuries that these deeds cover fall outside the scope of most digital knowledge infrastructure. This paper thus asks two questions central to entity-based archival description:

RQ1 What characterizes the long-tail population of person entities that provenance-based finding aids exclude, and how is that population distributed across the corpus?

RQ2 What relational structures among person entities can be recovered through typed relation extraction from formulaic notarial language, and what cross-volume and cross-notary connections does relation extraction expose that hierarchical archival description cannot represent?

We contribute (1) an empirical characterization of a long-tail population at corpus scale, demonstrating that NER extraction and authority file matching fail on opposite ends of the same naming spectrum, with direct consequences for Large Language Models (LLMs) and Knowledge Graphs (KGs), and (2) a typed relation extraction pipeline that, following the RiC standard's reconception of archival description around entities and their relationship graphs [1], produces structured output as a first step toward knowledge graph construction.

2. Related Work

This work bridges archival infrastructure theory, the computational challenges of long-tail data, and natural language processing (NLP) on historical texts. Archival description has been analyzed as infrastructure that structurally determines what can be discovered and what remains hidden within a collection [6]. Duff and Harris [2] showed that descriptive choices about what to foreground construct particular accounts of what records contain, often prioritizing institutional creators.

[1] "Draft Deeds", in Dutch

To mitigate this, emerging frameworks like the Records in Contexts (RiC) standard [1] reconceptualize hierarchical description: persons mentioned in records are granted independent descriptive status and can be connected to other persons through explicitly typed relationships, making entities searchable across separate volumes and fonds without prior knowledge of their names. Geerlings and Vriend [9] demonstrated the viability and benefit of implementing this entity-based description under RiC to early Dutch notarial records at the SAA. However, these implementations relied on manual extraction from finding aids that already contained rich, person-level information. On the other hand, when a finding aid completely omits such person-level information (as is the case with the provenance-bound descriptive entries of fonds 5075), person entities can only be recovered directly from the raw source texts. Scalably lifting these entities into archival view requires recognition of these entities with computational methods.

To recover these entities automatically, NLP techniques must adapt to the unique friction of historical language. Off-the-shelf Named Entity Recognition (NER) models trained on modern corpora have been found to perform poorly on historical texts due to non-standardized spelling, archaic vocabulary, and syntax that departs significantly from modern training data [10]. To counter this, domain adaptation is critical; for instance, applying NER with domain adaptation to a single notary's series of fonds 5075 within the Amsterdam City Archives yielded an adapted F1 of 0.735, demonstrating the value of localized training [11]. Concurrently, research shows that the highly repetitive, formulaic language characteristic of legal and notarial genres can effectively support rule-based extraction methods that remain highly competitive with statistical approaches [12]. These studies show that historical entity extraction benefits from localized domain adaptation and that the repetitive, formulaic language of legal and notarial genres can support effective rule-based extraction. What remains less understood is how these methods and their outputs behave when extended from individual series or selected samples to entire archival collections. At that scale, automated extraction makes visible a deeper structural feature inherent to historical demographic data: the massive, uneven distribution of the recovered population.

When historical populations are extracted from primary sources at scale, the resulting data often exhibit severe long-tail distributions. In computational linguistics, language models are known to fail disproportionately on entities residing outside the high-frequency head of a popularity distribution [13]. In historical domains, this data sparsity is compounded by historical preservation biases; traditional authority files and biographical databases disproportionately represent well-documented elites, leaving ordinary populations completely absent even when they appear repeatedly in primary records [14].

Furthermore, historical naming conventions frequently follow a *Large Number of Rare Events* distribution, where the vast majority of unique name strings appear only a single time across a corpus [15]. This makes traditional cross-referencing or deterministic record linkage unfeasible for the bulk of the population. While these studies identify the mathematical and systemic presence of the long tail, less is known about how these conditions interact with automated entity extraction pipelines and archival descriptive standards. In particular, the relationship between long-tail structure, NER performance, authority-file reconciliation, and emerging descriptive standards remains underexamined.

3. Methodology

Our pipeline adopts a two-phase strategy to confront archival populations' descriptive invisibility. The first phase (Sections 3.1–3.3) focuses on the extraction and empirical characterization of individual name strings to investigate where standard name-based indexing, entity extraction, and authority layers structurally fail. The second phase (Section 3.4) shifts from isolated strings to relational context, leveraging the formulaic syntax of early modern legal Dutch documents to anchor these long-tail individuals within a typed graph. All preprocessing, alignment, NER training and evaluation, population analysis, and relation extraction code, together with configuration files and scripts to reproduce the experiments in this section, are available in the project's GitHub repository.[2]

[2] https://github.com/tjude2003-spec/descriptive-invisibility

3.1. Corpus and Alignment

Three data sources about SAA's fonds 5075 were aligned by inventory number, the unique numerical identifier assigned to each physical notarial volume. The SAA finding aid provides 2,069 structured entries describing volumes by notary, document type, and date range. VeleHanden [4] crowdsourced annotations cover 298,684 deeds with 1.6 million person-name tags across 135 notaries. Transkribus [3] HTR transcriptions render the full text of digitized notarial volumes machine-readable. The final, aligned corpus comprises 2,282 inventories across 68 notaries. Of these, 339 inventories have both HTR transcriptions and VeleHanden annotations and form the evaluation subset for NER; the remaining 1,943 inventories form the extension corpus from which person names are extracted at scale.

3.2. NER Model Selection

Four NER models were evaluated on 73 held-out test inventories containing 66,616 unique ground-truth person names from VeleHanden. Three models were used off the shelf: (1) BERTje [16], a Dutch BERT-based transformer model; (2) spaCy's `nl_core_news_lg`[3] pipeline [17], a large Dutch model with pretrained word vectors; and (3) Flair's `ner-dutch-large`[4] tagger [18], based on contextual string embeddings. The fourth (4) model is a domain-adapted spaCy model, whose NER component was initialized from `nl_core_news_lg` and fine-tuned using distantly supervised labels derived from VeleHanden annotations. For each deed, VeleHanden person-name tags are fuzzy-matched against the HTR text using `RapidFuzz` ratio at a threshold of 0.88, selected after testing showed that lower values produced systematic false positives from short names and common particles while overall yield remained stable, to generate BIO-tagged training data. Training uses 204 inventories (184,996 sentences, 580,939 entity tokens) with strict cross-notary splits to prevent leakage between training and test notaries; the development set comprises 62 inventories (54,199 sentences). A manual audit of a sample of the resulting training entities classified approximately 65% as clearly correct matches, 18% as boundary errors, 13% as plausible but degraded by HTR noise, and 4% as genuine false positives. Evaluation is performed at the entity level, reporting macro-averaged F1 across inventories. Because the domain-adapted model (4) was the best performing, more than doubling the off-the-shelf spaCy baseline (see Section 4), it was used for all subsequent analyses.

3.3. Population Analysis

To map the characteristics of the long-tail population, the domain-adapted NER model was deployed across all 2,282 inventories, yielding 5,715,381 person-name mentions and 1,497,744 unique entity strings. Analyses requiring sampling used simple random draws without replacement, with fixed seeds recorded in the repository for reproducibility. The extracted population was subjected to an analytical pipeline consisting of the following computational steps:

1. **Frequency Distribution:** To establish whether the extracted population follows the Large Number of Rare Events pattern characteristic of historical naming [15], string recurrence was measured across inventories using exact matching on lowercased text blocks. This quantifies the structural sparsity of the data. No automated entity resolution or fuzzy merging was applied at this stage, establishing unique strings rather than individuals as the baseline unit of analysis.
2. **Name Parsing:** This analysis measures the amount of identifying information available in each name string, which determines both how well NER extracts the name string and whether it can be matched to an authority file. Each unique entity string was decomposed into typed components (*given name*, *patronymic*, *prefix*, *family name*) using a rule-based regular expression parser. Patronymics and gender markers were isolated by matching against early modern Dutch patronymic suffixes, specifically targeting variations of *-sz*, *-sen*, *-zoon*, *-dr.*, and *-dochter*. Based on

[3] https://huggingface.co/spacy/nl_core_news_lg/tree/main
[4] https://github.com/flairnlp/flair

the extracted components, strings were classified into one of six structural naming configurations to measure identifying information density.

3. **Orthographic Variation:** This analysis estimates the extent to which historical spelling variation and HTR corruption fragment the extracted population into string-level duplicates that resist automated merging. A random sample of 10,000 entity strings was drawn from the full extracted population. Pairwise string comparison was executed across these multi-token strings to quantify spelling fragmentation and HTR corruption. Token similarity was computed using the `token_sort_ratio` metric from the `RapidFuzz` library. String pairs returning a score of ≥ 0.85 were flagged as orthographic variants and connected into clusters via a connected components algorithm. To assess the reliability of the threshold, all flagged pairs were manually classified into three categories: genuine spelling variants (shared given name and family name), given-name-only overlap (different family name), and coincidental similarity.
4. **Authority Alignment:** This analysis tests whether extracted entities can be anchored to external knowledge bases. Candidate matching was performed against two external authority layers using the parsed fields generated in step 2. A sample of 10,000 extracted strings, to maintain consistency with the orthographic variation analysis, was aligned against the SAA's 1,917 expert-reconciled *Person Reconstructions* (structured using the ROAR ontology framework) and the ECARTICO prosopographical database ($> 55{,}000$ records) [5]. Candidate matches were evaluated using exact matching on the combined *given name* and *family name* fields.[5]

Further details, including matching and threshold configurations and execution scripts for these pipeline steps, are documented in the project GitHub repository.

3.4. Typed Relation Extraction and Network Construction

Because notarial deeds encode familial relationships in formulaic language [19], we implement a rule-based typed relation extraction framework adapted from the MiSS (Mining Social Structures) architecture [20, 15] to bypass the limitations of thin name strings by linking entities contextually. The pipeline processes sequential NER outputs page-by-page, examining both co-occurring entity pairs and pairs separated by exactly one intervening string (to capture instances containing an intervening witness or co-signer). The textual span between target entities is capped at 150 characters to exclude coincidental co-occurrences and matched against a curated dictionary of 81 indicator phrases mapping to five familial categories: *spouse*, *child*, *sibling*, *widow*, and *widower*. We expanded MiSS's 26 published indicators [20, 15] to 81 through iterative corpus review: unmatched between-entity text spans were sorted by frequency, classified as relational indicators or noise, and added to the dictionary until successive passes yielded no new types. The dictionary reactivates child relation markers that MiSS defines but does not use, adds sibling indicators, and incorporates historical spelling variants and scribal abbreviations.[6] To handle HTR noise while maintaining high precision, the system records a match if the span matches an indicator exactly or falls within a ± 3 character prefix/suffix slack bound. For *spouse* relations, a deterministic gender filter automatically rejects same-gender pairs by evaluating gendered patronymic suffixes (e.g., masculine *-sz* vs. feminine *-dr.*)[7] while strings lacking explicit suffixes pass by default. Validated matches are output as structured triplets (*person_1*, *relation_type*, *person_2*) anchored to their specific inventory, page, and text span, conforming with the Records in Contexts (RiC) standard.

Triplets are compiled into a network graph using `NetworkX`, with unique name strings as nodes and relations as edges, allowing the extraction of relational structures across distinct physical volumes. To separate legitimate orthographic variation from artificial clusters generated by patronymic collisions (e.g., merging unrelated individuals under "jan jansz"), components with ≥ 3 nodes must satisfy a two-step constraint to be flagged as collapsible variant groups: (1) given name consistency, where

[5] https://github.com/tjude2003-spec/descriptive-invisibility/blob/main/analysis/external_recognition_decomp.py
[6] https://github.com/tjude2003-spec/descriptive-invisibility/blob/main/relations/build_relations_fast.py
[7] Reflecting the legal constraints of the early modern Netherlands, where same-sex marriage was unrecognized until its legalization in 2001.

unique given names constitute $\leq 50\%$ of the cluster's strings, and (2) degree thresholding, placed at the largest gap in the component's node degree distribution to isolate tight variant sets from sprawling hubs. Nodes remain defined as raw name strings rather than resolved historical individuals. Finally, we isolate a single notary sub-network for qualitative validation as a cross-volume case study: Jan Verleij (83 inventories, aligning with baseline historical NLP precedents [11]).

4. Results

4.1. NER Model Selection

To address the preliminary question of whether robust NER is possible on this material at scale, we first assessed how well contemporary Dutch NER models transfer to early modern notarial text. Table 1 reports macro-averaged precision, recall, and F1 scores at the entity level on 73 held-out test inventories. The four-model comparison confirms a substantial domain mismatch between modern NER training corpora and historical notarial language: off-the-shelf models achieve macro F1 in the 0.16–0.31 range. BERTje performs worst (F1 = 0.164), consistent with its wordpiece tokenizer fragmenting unfamiliar historical forms. The domain-adapted spaCy model more than doubles F1 relative to the off-the-shelf spaCy baseline (from 0.302 to 0.680) and raises precision from 0.210 to 0.703. Error analysis indicates that much of this precision gain comes from learning to reject formulaic legal terms (*comparant*, *voornoemde*) that occupy syntactic positions where the model initially expected person names.

Model	Precision	Recall	F1
BERTje (off-the-shelf)	0.127	0.263	0.164
spaCy `nl_core_news_lg` (off-the-shelf)	0.210	0.584	0.302
Flair `ner-dutch-large` (off-the-shelf)	0.337	0.740	0.449
spaCy (domain-adapted)	0.703	0.696	0.680

Table 1
Entity-level NER performance on historical notarial text, macro-averaged across inventories.

Importantly, because the VeleHanden crowdsourced coverage is inherently incomplete, the evaluation framework penalizes the adapted model: any genuine person name extracted by the model that was missed by volunteers is strictly counted as a false positive. To quantify this effect, we conducted a manual audit of a random sample of 100 false positives generated by the adapted model. The analysis revealed that 64% were actually genuine person names absent from the ground-truth index, 34% were partial name extractions, 1% was ambiguous, and only 1% constituted a true false positive (a non-person string). Adjusting for this ground-truth under-coverage suggests that the adapted model's true precision is likely substantially higher than reported, establishing the domain-adapted spaCy model as a sufficiently reliable instrument for the corpus-scale extractions that follow.

4.2. Characterizing the Long-Tail Population

Frequency structure Of 1,497,744 unique entity strings, 82.0% appear in exactly one inventory and 96.5% in five or fewer (Figure 1). The twenty most frequently recurring strings are all patronymic forms: "jan jansz" (1,529 inventories), "jan pietersz" (1,325), "pieter jansz" (1,192). These high-frequency strings do not represent highly visible individuals but names so structurally common that many unrelated people share them. Because the early modern Dutch patronymic system generates identical strings for anyone sharing a common given-name and father's-name combination (e.g., "Jan Jansz" translates to "Jan, son of Jan"), frequency in this corpus signals structural naming conventions rather than individual historical prominence.

Name composition Parsing all extracted entity strings into typed components reveals three dominant structures: given name + family name (67.1%), given name + prefix + family name (18.3%), and given

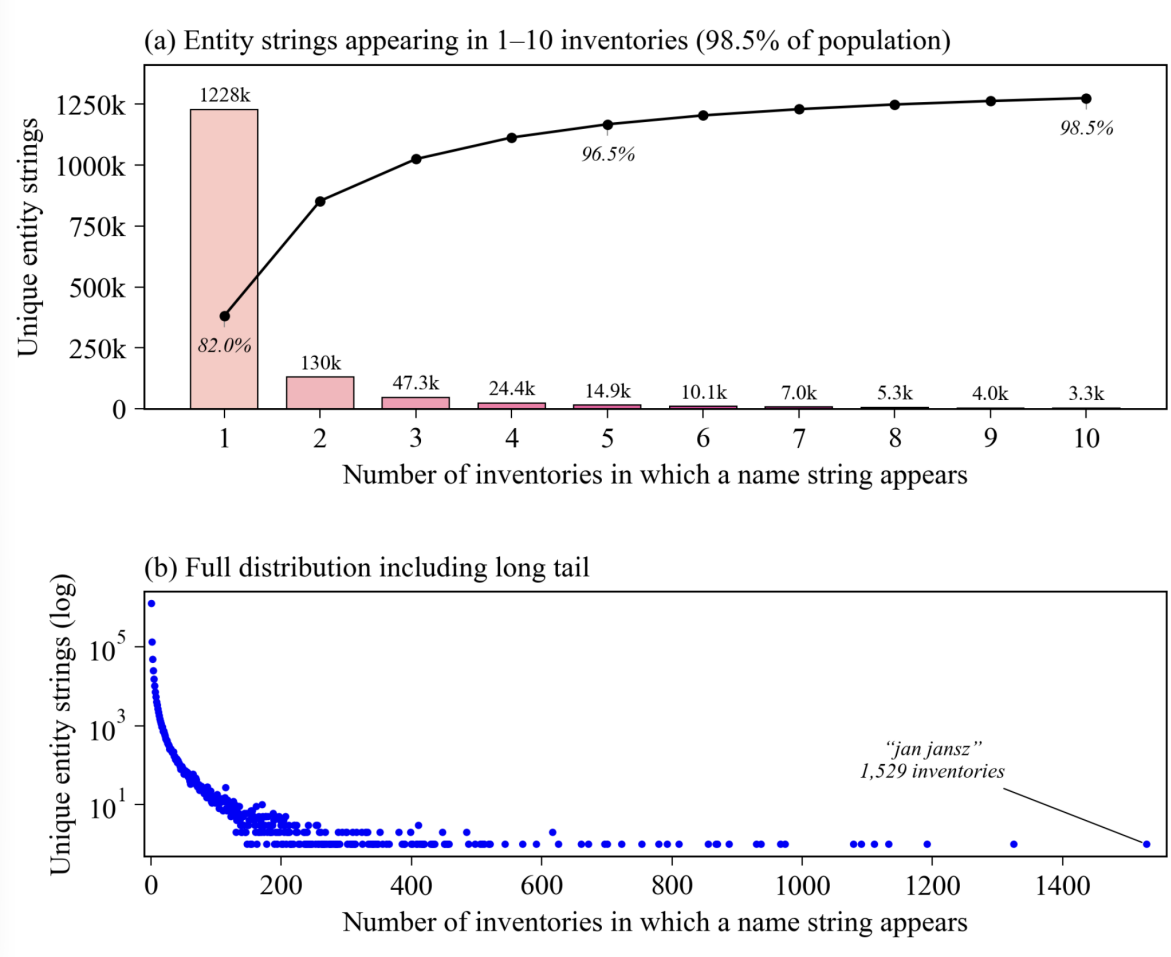


Figure 1: Distribution of entity strings by number of inventories in which they appear. 82% of strings appear in exactly one inventory.

name + patronymic (7.8%). Together, these configurations account for 93.2% of the corpus. Crucially, names carrying both a patronymic and a family name–the structural name composition providing the densest identifying information for entity disambiguation–account for only 5.6%.

Orthographic variation At a 0.85 similarity threshold, 13.0% of the strings possessed at least one near-duplicate. Manual auditing of these pairs revealed that 35.6% were genuine spelling variants sharing both given and family names, 50.5% shared a given name but differed fundamentally in their family name, and 13.9% were purely coincidental lexical similarities. While pair-level clusters successfully isolated plausible variant groupings (e.g., {"elisabet de fervaques," "elisabeth de ferraecques"}), larger multi-element clusters rapidly degraded because a single false positive match acts as a bridge, pulling entirely unrelated names into the chain. Consequently, string similarity alone cannot reliably reconcile variants in a population where thousands of distinct individuals share common given names.

Candidate matching: extractability–matchability opposition The aforementioned naming structures interact antagonistically with downstream authority file matching pipelines. Cross-referencing the full extracted entity corpus against two authority layers produced minimal overlap. Out of the SAA's 1,917 Person Reconstructions, only 3 matched entities in a 10,000-entity sample (0.03%), while ECARTICO yielded 242 candidate matches (2.4%). More important than the low matching rate is how matching success diverges by name structure. A given name paired strictly with a patronymic produces zero ECARTICO candidates because the database indexes by family name; individuals without one

remain unmatchable regardless of database size. Adding a family name raises the match rate to 8.2%.

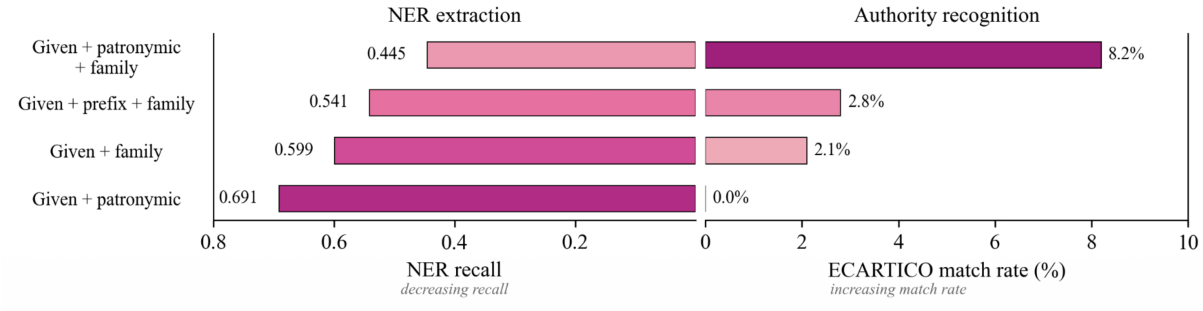


Figure 2: NER recall (left) and ECARTICO match rate (right) by name structure type. Recall is highest for patronymic names; authority file matching is highest for names carrying family names.

NER recall runs in the opposite direction across the same naming structures (Figure 2). Recall is 0.691 for short patronymic names and drops to 0.445 for longer names carrying family names. What NER extracts most reliably, authority databases cannot match. What authority files could anchor, NER extracts least reliably. Short names survive HTR corruption because fewer tokens can be degraded, but they lack the family-name component that authority files index by. Long names carry that component but accumulate more recognition errors across their longer spans.

4.3. Typed Relation Extraction

The relation extraction procedure identified 44,249 typed relations across 2,194 of the 2,282 inventories (96%). Widow relations dominate the dataset (28,946), followed by spouse (13,454), child (1,655), sibling (192), and widower (2). A stratified sample of 794 relations (200 widow, 200 spouse, 200 child, 192 sibling, 2 widower) was manually validated, achieving 100% precision. This perfect precision is a direct artifact of our strict between-entity rule-based matching constraints, which rely on highly conserved, deterministic legal formulas. However, this precision comes at the expense of recall: given the domain-adapted NER model's baseline recall of 0.696, the theoretical probability of successfully extracting both entities required to form a relation pair is bounded at approximately 0.48 (0.696×0.696), creating a strict performance ceiling before the relation extraction logic is applied.

Network structure The 44,249 relations map onto a network of 62,111 unique name strings through 41,791 edges. The network is highly fragmented: 80% of clusters are dyads. The largest connected component spans 10,584 strings (17% of nodes), produced by patronymic collision—common strings like "jan jansz" (352 connections) act as false hubs welding unrelated families. Of 4,254 clusters with three or more nodes, 490 are legitimate orthographic-variation groups based on the $\leq 50\%$ given-name consistency filter.

Cross-archival-unit connections Of the 41,791 distinct edges in the network, 1,651 link name strings appearing in different inventories, and 242 of those cross different notaries. The cross-inventory signal is heavily concentrated in distinctive family-name forms rather than common patronymics: only 7 of the 1,651 cross-inventory edges connect nodes where both entities carry exclusively patronymic names. These edges expose latent relational structures that traditional hierarchical archival description, organized by notary, has no structural mechanism to represent.

Case study: Jan Verleij The sub-network generated from the 83 inventories of notary Jan Verleij contains 2,082 nodes and 1,163 relations, featuring a prominent cluster of six entity strings linked through a central "pieter jansz" node (Figure 3). Because this cluster contains four distinct given names across its six strings, standard name-parsing heuristics would discard it as an unresolvable patronymic

collision. However, the extracted typed relations reconstruct a coherent family structure scattered across five separate inventories: three child relations (indicated by *soon van*"[8]) identify "pieter jansz" as the son of three distinct spelling variants of the same individual ("jan everts", "jan evertse", and "jan evertsz"), a spouse relation connects "maria ter meetelen" to "pieter jansz" as his wife, and a widow relation connects "sara de gaert" to "jan everts", identifying her as Jan's widow. Rather than a random collision of strings, the network maps out a father (Jan, in three spellings), his son (Pieter), whose patronymic "jansz" independently confirms the filiation, Pieter's wife (Maria), and Jan's widow (Sara): four individuals across two generations, illustrating the utility of entity-based description under the RiC standard, where long-tail historical actors become discoverable through their explicit cross-document social relationships rather than requiring prior, exact knowledge of their names.

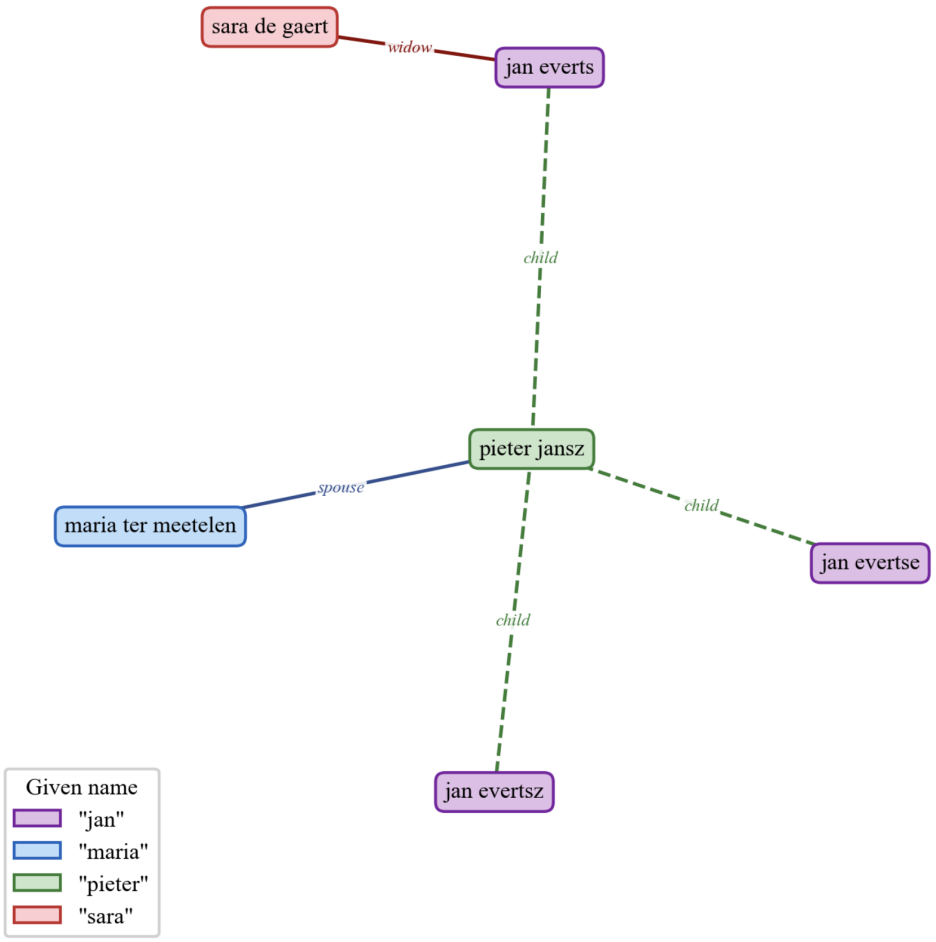


Figure 3: Sub-network cluster from notary Jan Verleij showing a family reconstructed through typed relation extraction across five inventories. Child relations point from child to parent.

5. Discussion and Conclusion

The results validate our conceptualization of descriptive invisibility as distinct from archival silences [7, 8]. The recovery of 1.49 million unique strings and 44,249 typed relations proves that this population was never omitted from the archive; it was fully preserved but rendered unreachable by the access infrastructure. The finding aid, computational extraction, and external authority files each operate according to their own organizational logic, and each independently leaves the same population unreachable. Provenance-based description excludes all persons because its schema holds no person-name category. NER extraction recovers those persons but does so unevenly; short patronymic names

[8] son of

are extracted reliably because fewer tokens are exposed to HTR corruption, while structurally rich names are lost to recognition errors. Authority files could anchor those structurally rich names, but those are the names extraction delivers least reliably. This opposition follows directly from the historical Dutch naming convention, not from any specific model or database, and becomes visible only when the layers are read together because each, examined in isolation, operates precisely as designed.

Early modern Dutch patronymics and frequent name collisions frustrate entity linking and authority reconciliation, but formulaic notarial language offers another route to structuring archival populations. Our typed relation extraction makes individuals searchable through localized social networks even where their names lack authority-file matches. The Jan Verleij case shows how multi-generational relations spanning five inventories can distinguish variant spellings that string comparison would treat as unrelated or conflate, making relational structure a high-precision identity-resolution heuristic.

This work illustrates the utility of RiC's entity-based description: long-tail actors become discoverable through their explicit cross-document social relationships rather than requiring prior, exact knowledge of their names. Whereas Geerlings and Vriend [9] manually extracted 59 agents from a single fonds, and VeleHanden crowdsourced indexes for 298,684 deeds, neither approach can reconcile 1.5 million unique strings at corpus scale. The pipeline in this work therefore provides structured starting material for knowledge-graph construction, directing rather than replacing human judgement in bringing long-tail notarial populations into view.

Limitations Name strings are not persons: spelling variations and patronymic collisions can distort population statistics. The relation extraction pipeline's theoretical recall ceiling of approximately 0.48 (0.696×0.696) compounds with the rule-based matching constraints to limit coverage. The 413:1 ratio of male to female patronymic markers conflates naming-convention properties with potential NER underdetection of female actors; the gender-suffix parser treats linguistic suffixes as proxies for gender without independent validation, and the extreme ratio may partly reflect the parser's reliance on a limited set of suffix patterns rather than the historical record alone. All manual classification tasks reported in this paper—the false-positive audit, orthographic variant annotation, and relation validation—were performed by the first author; no inter-annotator agreement was computed. Finally, relying on VeleHanden annotations for distant supervision and evaluation introduces a circular dependency, which our false-positive analysis partially mitigates but does not fully eliminate.

Future work To transition from raw name strings toward identified historical persons, future work could implement collective entity resolution frameworks that leverage relational network context [21]. Conducting gender-stratified NER evaluation would separate naming-convention effects from model bias. The distant supervision strategy could be extended by combining VeleHanden name annotations with the 81 typed relation indicators to generate silver training data for end-to-end relation extraction models such as GliREL [22], potentially recovering relations that the current rule-based approach misses. Deploying this pipeline across international archives with divergent provenance traditions would help establish whether descriptive invisibility is a general feature of hierarchical, creator-based archival description or a condition specific to Dutch notarial holdings.

Acknowledgments

We thank Merel Geerlings and Pauline van den Heuvel at SAA for sharing expertise on entity-based description and access to the Person Reconstructions data, and Ivo Zandhuis for consultation on relation extraction and ECARTICO. The VeleHanden volunteers who have indexed notarial deeds since 2018 made both the training data and the evaluation framework for this work possible.

Generative AI Declaration

Google Colab's Gemini was used to debug code and for grammar check. All primary textual content, analysis, and visualizations were produced by the authors.

References

[1] ICA Expert Group on Archival Description, Records in contexts: A conceptual model for archival description (RiC-CM v0.2), International Council on Archives, 2021. URL: https://www.ica.org/ica-network/expert-groups/egad/records-in-contexts-ric/.

[2] W. M. Duff, V. Harris, Stories and names: Archival description as narrating records and constructing meanings, Archival Science 2 (2002) 263–285. doi:10.1007/BF02435625.

[3] G. Muehlberger, L. Seaward, M. Terras, S. Ares Oliveira, V. Bosch, M. Bruber, et al., Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study, Journal of Documentation 75 (2019) 954–976. doi:10.1108/JD-07-2018-0114.

[4] VeleHanden, Crowdsourcing platform for cultural heritage, n.d. URL: https://velehanden.nl/, accessed 2026.

[5] M. J. Bok, H. Nijboer, ECARTICO: Linking cultural industries in the early modern Low Countries, ca. 1475–ca. 1725, University of Amsterdam, n.d. URL: http://www.vondel.humanities.uva.nl/ecartico/.

[6] G. C. Bowker, S. L. Star, Sorting Things Out: Classification and Its Consequences, MIT Press, 2000.

[7] R. G. S. Carter, Of things said and unsaid: Power, archival silences, and power in silence, Archivaria 61 (2006) 215–233. URL: https://archivaria.ca/index.php/archivaria/article/view/12541.

[8] M.-R. Trouillot, Silencing the Past: Power and the Production of History, Beacon Press, 1995.

[9] M. Geerlings, N. Vriend, Proof of contexts: New accessibility for archive users, Amsterdam City Archives & Noord-Holland Archives, 2024. Final report.

[10] M. Ehrmann, A. Hamdi, E. L. Pontes, M. Romanello, A. Doucet, Named entity recognition and classification on historical documents: A survey, ACM Computing Surveys 56 (2023) 1–47. doi:10.1145/3604931.

[11] B. Hendriks, P. Groth, M. van Erp, Recognising and linking entities in old Dutch text: A case study on VOC notary records, in: Proceedings of the International Conference Collect and Connect, 2020. URL: http://ceur-ws.org/Vol-2810/paper3.pdf.

[12] M. Koolen, R. Hoekstra, Detecting formulaic language use in historical administrative corpora, in: Proceedings of the Computational Humanities Research Conference (CHR 2022), CEUR-WS.org, 2022, pp. 127–151. URL: https://ceur-ws.org/Vol-3290/long_paper5740.pdf.

[13] A. Mallen, A. Asai, V. Zhong, R. Das, D. Khashabi, H. Hajishirzi, When not to trust language models: Investigating effectiveness of parametric and non-parametric memories, in: Proceedings of ACL 2023, 2023, pp. 9802–9822. doi:10.18653/v1/2023.acl-long.546.

[14] M. Ripoll, N. Reeves, A. Kurteva, E. Simperl, A. Meroño Peñuela, K. Diepold, Filling in the blanks? A systematic review and theoretical conceptualisation for measuring Wikidata content gaps, arXiv preprint arXiv:2505.16383 (2025). URL: https://arxiv.org/abs/2505.16383.

[15] J. Efremova, A. Montes García, A. Bolt Iriondo, T. Calders, Who are my ancestors? Retrieving family relationships from historical texts, in: Information Retrieval (RuSSIR 2015), volume 573 of *Communications in Computer and Information Science*, Springer, 2016, pp. 121–129. doi:10.1007/978-3-319-41718-9_6.

[16] W. De Vries, A. van Cranenburgh, A. Bisazza, T. Caselli, G. van Noord, M. Nissim, Bertje: A dutch bert model, arXiv preprint arXiv:1912.09582 (2019).

[17] M. Honnibal, spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing, (No Title) (2017).

[18] A. Akbik, T. Bergmann, D. Blythe, K. Rasul, S. Schweter, R. Vollgraf, Flair: An easy-to-use framework for state-of-the-art nlp, in: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics (demonstrations), 2019, pp. 54–59.

[19] L. Nussdorfer, Brokers of Public Trust: Notaries in Early Modern Rome, Johns Hopkins University Press, 2009.
[20] B. Ranjbar-Sahraei, J. Efremova, MiSS: Mining social structures, n.d. URL: https://github.com/branjbar/MiSS-Information-Extraction, software.
[21] I. Bhattacharya, L. Getoor, Collective entity resolution in relational data, ACM Transactions on Knowledge Discovery from Data 1 (2007) Article 5. doi:10.1145/1217299.1217304.
[22] J. Boylan, C. Hokamp, D. G. Ghalandari, Glirel – generalist model for zero-shot relation extraction, 2025. URL: https://arxiv.org/abs/2501.03172. arXiv:2501.03172.