Linking Historical Persons to Archival Documents: Challenges and Approaches*

Huan Chen[1,2,*,†], Gareth J. F. Jones[2,3,†], Declan O'Sullivan[2,4,†], Eamonn Kenny[2,4,†], Alex Randles[2,4,†], Neil Johnston[5,†] and Rob Brennan[1,2,†]

[1]School of Computer Science, University College Dublin, Belfield, Dublin 4, Ireland
[2]ADAPT Centre, O'Reilly Institute, Trinity College Dublin, Dublin 2, Ireland
[3]Hamilton Institute, Maynooth University, Maynooth, Co. Kildare, Ireland
[4]School of Computer Science and Statistics, Trinity College Dublin, Dublin 2, Ireland
[5]The National Archives, Richmond TW9 4DU, United Kingdom

Abstract

The State Papers Ireland (TNA SP 60- SP 63) constitute a rich and complex historical corpus, including official letters, private papers, petitions, and correspondence preserved at The National Archives UK, with a current focus on 1660–1715. Linking person mentions in these records to entities in the Virtual Record Treasury of Ireland (VRTI) Knowledge Graph (KG) is challenging due to historical spelling variants, temporal uncertainty, and incomplete metadata. This paper investigates methods for linking archival documents to their corresponding historical individuals, which are recorded in the VRTI-KG. We propose a metadata-driven heuristic approach—utilising fuzzy string matching and temporal constraints, and compare it against a neural entity linking method that employs BERT-based embeddings. Our results indicate that while neural models offer semantic flexibility, a metadata-driven heuristic method remains more robust for historical corpora characterised by sparse and inconsistent metadata, spelling variation, and heterogeneous location information. To further enhance linking accuracy, we discuss the potential of Named Entity Recognition (NER) to enrich both entity properties and document features, providing additional cues for mapping archival documents to the corresponding KG person entities.

Keywords

Entity Linking, Historical Documents, Knowledge Graphs, Digital Humanities, Name Disambiguation

1. Introduction

Historical documents serve as the primary source of information for understanding cultural evolution [1]. However, there are substantial obstacles to scholarship due to the large amount and frequently damaged nature of archive documents [2]. Digital Humanities (DH) research addresses these challenges by providing frameworks for organising and computationally analysing archival corpora [3]. The Virtual Record Treasury of Ireland (VRTI)[1] has developed a Linked Open Data Knowledge Graph (KG) alongside archival scans and metadata to digitally reconstruct Ireland's national archive destroyed in 1922. A core corpus of this reconstruction is the State Papers Ireland (1660–1715)[2], which for now comprises official correspondence and administrative records from the Restoration to the end of Queen Anne's reign [4]. These records provide foundational evidence for early modern governance and communication between

SemDH 2026: Third International Workshop of Semantic Digital Humanities. Co-located with ESWC 2026, May 10, 2026, Dubrovnik, Croatia

*Corresponding author.
†These authors contributed equally.
huan.chen2@ucdconnect.ie (H. Chen); Gareth.Jones@mu.ie (G. J. F. Jones); Declan.OSullivan@tcd.ie (D. O'Sullivan); eamonn.kenny@adaptcentre.ie (E. Kenny); alex.randles@adaptcentre.ie (A. Randles); Neil.Johnston@nationalarchives.gov.uk (N. Johnston); rob.brennan@ucd.ie (R. Brennan)
0009-0007-8549-0414 (H. Chen); 0000-0003-2923-8365 (G. J. F. Jones); 0000-0003-1090-3548 (D. O'Sullivan); 0000-0003-1895-1368 (E. Kenny); 0000-0001-6231-3801 (A. Randles); 0000-0002-6145-9611 (N. Johnston); 0000-0001-8236-362X (R. Brennan)

© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

[1]https://virtualtreasury.ie/knowledge-graph
[2]https://virtualtreasury.ie/gold-seams/state-papers-ireland

English administrations in Dublin and London; these records are indispensable for reconstructing Ireland's administrative past following the loss of the Public Record Office of Ireland [5].

Enabling meaningful prosopographical research, network analysis, and structured querying requires linking the persons referenced in these documents to unique person identifiers in the KG [6]. Within the VRTI KG, each manuscript is distilled into structured metadata (including fields such as sender, recipient, and issued at [location]) to provide a foundational context for this process. However, assigning unique KG person identifiers to manuscript metadata remains a significant challenge[1]. Individuals often appear under multiple historical name or title spelling variants, distinct individuals may share identical names, and contextual evidence is frequently sparse or ambiguous. Furthermore, the contents of structured metadata fields such as "sender", "recipient", "issued at", and "sent to" do not always directly correspond to biographical information such as birth or death places. Many relevant historical features, such as social roles, kinship ties, military ranks, and institutional affiliations, are not encoded in the metadata.

Research Question: How can historical documents be accurately linked to structured person entities in a KG despite ambiguous names, temporal uncertainty, and incomplete or inconsistent metadata and context?

To address this question, we evaluate a metadata-driven heuristic method combining fuzzy name matching with temporal constraints and location features, and we also explore a neural entity linking method to assess whether contextual embedding representations improve disambiguation performance in a historical domain. Our code and sample data is on GitHub[3].

The main contributions of this work are:

- We formalise the task of linking archival documents, such as the State Papers Ireland documents, to structured person entities in a KG using available archival metadata and KG properties and concepts.
- We propose and evaluate a metadata-driven heuristic entity linking approach combining fuzzy name matching with temporal and location constraints, enabling explicit feature weighting to support domain-expert control in DH settings.
- We compare this heuristic approach with a neural entity linking method and analyse their limitations in a historical domain with spelling variation and sparse contextual information.
- We provide an empirical assessment of the challenges inherent in person entity disambiguation within early modern archival collections and discuss strategies for property-rich Named Entity Recognition to enrich KGs linked to archival documents.

The remainder of this paper is organised as follows: §2 presents the problem statement for archival document to person linking, §3 describes the use cases to be addressed, and §4 reviews related work. §5 details the methodology, while §6 outlines the experimental setup and results. Finally, §7 provides a discussion, and §8 concludes the paper and suggests directions for future work.

2. Problem Statement

The core objective is to solve the **Archival Document-to-Person Linking task**: resolving mentions of individuals in the archival metadata (derived from transcription of scanned historical documents) of the State Papers Ireland to unique, persistent identities within the VRTI KG State Papers people.

The KG used in this work is structured according to the VRTI ontology[4], which provides semantic definitions for archival entities and their properties. Person entities are represented using properties such as canonical and alternative names (e.g., `vrti`[5]`:Name`), temporal attributes including birth and death dates or floruit intervals, and event-based biographical modelling derived from CIDOC-CRM[7]. For example, birth events (`crm`[6]`:E67_Birth`) are linked to persons via `crm:P98_brought_into_life`

[3]https://github.com/Foina/metadata-driven-heuristic
[4]https://ont.virtualtreasury.ie/ontology/index-en.html
[5]<https://www.w3id.org/virtual-treasury/ontology#>
[6]<http://erlangen-crm.org/current/>

and associated with places using `crm:P7_took_place_at`.

The State Papers Ireland items consist of correspondence, petitions, and administrative records from 1660–1715. Each item is accompanied by manuscript-level metadata extracted from the manually compiled archival catalogues. Key fields include:

- Title - summaries of the document content
- Date of creation - the document's creation date
- Sender - the individual or office responsible for the document
- Recipient - the intended recipient
- Issued at - the location from which the document was sent
- Sent to - the location to which the document was sent
- Scope and content - descriptive summaries of the content

These metadata fields provide the input for the linking process. However, historical practices and transcription limitations mean that metadata fields may be missing, inconsistent, or ambiguous: names may appear under multiple historical variants, locations may differ from KG records, and dates may be partially specified. Early Modern spelling is extremely irregular and can differ significantly from modern spelling. Consequently, the sparse, inconsistent, and ambiguous nature of manual archival metadata makes reliable identification of historical individuals a challenging task.

To clarify how each metadata field corresponds to the KG representation of a person, Table 1 presents the correspondence between the document-level fields and person-related KG properties.

Metadata	Person-related KG Properties
Person's Name Fields	URI
Sender	Name
Recipient	Gender
Location-related Fields	
Issued_at	Birth place
Sent_to	Death Place
Temporal-related Fields	
Date_of_Creation	Era
Content_Date_Range	Birth Date
	Death Date
Text-related Fields	
Title	DIB
Scope_and_Content	Wikidata

Table 1
Correspondence between archival metadata fields and VRTI KG person properties

The task being executed may be summarised as a form of a query:
Query: Document metadata fields (Sender, Recipient, Date of creation, Issued at, Sent to).
Candidate sets: Person entities in the KG, described using properties such as `vrti:Name`, `vrti:Surname`, `vrti:VRTI_ERA`, `vrti:Floruit` and event-based relations (e.g., birth events linked to places via `crm:P7_took_place_at`).
Output: A ranked list of candidate person entities for each sender or recipient.

3. Use Cases

This section describes representative challenges in linking historical archival documents to person entities in historical KGs such as the VRTI KG.

- Use Case 1: Ambiguous Senders and Name Variants A central challenge in historical person linking concerns variation in recorded personal names across archival sources.

  - Example: Edmund Ludlow[7]
    * Name Variants:
      · Edmund Ludlow
      · Edward Ludlow
      · Simon of the Holy Spirit Ludlow
    * Birth/Death: 1616-1692
    * Place: Switzerland , Vevey

    All variants refer to the same individual and share identical life dates and associated external identifiers. However, different historical sources and archival descriptions record different forms of his name. A researcher examining a State Papers Ireland document may encounter a sender recorded as "Edward Ludlow" or "Edmund Ludlow". Without awareness of historical spelling variation or alternative name forms, linking this metadata entry to the correct KG entity is non-trivial.
    From a technical perspective, SPARQL queries alone are insufficient. SPARQL primarily supports exact or simple string matching, so variants such as "Edward Ludlow" and "Edmund Ludlow" will not match unless all aliases are explicitly enumerated. Moreover, SPARQL does not provide a natural mechanism for relevance ranking[8]. Consequently, additional reasoning or similarity assessment required for resolving historical ambiguities in this entity linking tasks.

- Use Case 2: Name, Temporal and Location Overlap
  Ambiguity also arises when multiple distinct individuals share the same canonical name and overlapping life spans.
  - Example: Two distinct KG entities (Richard Boyle[8], Richard Boyle[9]) share the same canonical name and their birth/death or floruit or active periods overlap in the Early Modern.
    * Richard Boyle (1566-1643); Kent, England; Cork, Ireland
    * Richard Boyle (1612-1698); Cork, Ireland; Yorkshire, England

    Both individuals are recorded as active in the seventeenth century (c.1600s), and their names appear identically in document metadata. When a State Papers Ireland record lists "Boyle; Richard" as a sender or recipient with only a creation year, it is difficult to determine which KG entity the document refers to. Straightforward retrieval or filtering using SPARQL or Python-based heuristics may return both candidates with similar scores.

- Use Case 3: Discrepancies Between Document Metadata Event Locations and KG locations
  State Papers Ireland metadata may include fields "Issued at" or "Sent to" that do not correspond to a person's birth/death place recorded in the KG.
  - Example: "Colonel and Governor William Leigh to General Ludlow at Duncannon fort, Waterford" [10]
    * Issued at: Waterford, Ireland
    * Sent to: Duncannon, Ireland
    * Recipient (KG entity): Edmund Ludlow
    * Recorded Place of Death in KG: Vevey, Switzerland

    The document records administrative event locations (Waterford and Duncannon), while the KG records life-event locations (e.g., the death place in Switzerland). These represent fundamentally different semantic categories: document metadata describes where communication occurred, whereas KG properties describe biographical life events(birth/death place). This semantic mismatch can mislead linking algorithms: a candidate may be incorrectly deprioritised if only biographical locations are considered.

[7] https://kg.virtualtreasury.ie/entitycard/person/Ludlow_Edmund_c17/v12d6sy
[8] https://kg.virtualtreasury.ie/entity-card/person/Boyle_Richard_c17/v1d57md
[9] https://kg.virtualtreasury.ie/entity-card/person/Boyle_Richard_c17/v15s7bb
[10] https://virtualtreasury.ie/item/TNA-SP-63-303-3

4. Related Work

This section provides an overview of the evolution of entity linking methods, with a focus on both neural approaches and their application to historical and cultural heritage corpora.

Entity linking standard approaches often involve two steps: candidate generation and candidate ranking [9]. Neural methods leverage high-dimensional embeddings to represent both mention and candidate entities [10, 11]. State-of-the-art models such as BLINK [11], GENRE [12], and CHOLAN [13] use transformer-based encoders to capture contextual information and rank candidate entities. Recent LLM-based methods, such as LLMAEL pipeline [14], augment candidate descriptions or mention representations with LLM knowledge to improve disambiguation, particularly for low-frequency or long-tail entities. However, historical and cultural heritage corpora present unique challenges for entity linking, including orthographic variation, sparse metadata, and ambiguous temporal or geographic information [1]. Benkhedda et al. [15] evaluated state-of-the-art neural models on archival content, noting that while embedding-based models excel on modern data like Wikipedia, they often struggle with the sparse context found in historical metadata.

Parallel research in large language model-based question answering has demonstrated the benefits of integrating NER features with KG constraints, for example, in Retrieval-Augmented Generation (RAG) [16] frameworks [17, 18]. These approaches show how structured entity and contextual information can enhance performance on tasks involving sparse or ambiguous data, analogous to the challenges in linking archival metadata to historical KG person entities.

To address similar limitations in historical corpora, NER is employed to identify and classify mentions of entities(e.g, persons, places, organisations) within the text [19]. In historical corpora, NER plays a crucial role in extracting structured information from archival documents, providing the foundational data required for linking documents to a KG [20]. Prior research has explored adapting NER models to historical texts using rule-based approaches, gazetteers, or machine learning models fine-tuned on historical corpora [1]. High-precision NER in historical corpora often relies on combining lexical heuristics with contextual cues to handle spelling variation and ambiguous mentions [21]. Ultimately, integrating these NER-extracted features with KG-based constraints offers a robust path forward, allowing for the enrichment of sparse archival metadata with the specific social roles and temporal data necessary for accurate historical disambiguation.

Despite strong progress in neural and KG-based entity linking methods, most existing approaches rely on complex, black-box architectures. While effective on modern datasets, they typically provide limited interpretability and little opportunity for domain experts to directly control the matching process. This limitation is particularly critical in Digital Humanities settings, where transparency, traceability, and expert-driven reasoning are essential for sparse and noisy historical metadata. This motivates lightweight heuristic methods with explicit feature weighting, which prioritise interpretability and expert control over purely data-driven optimisation.

5. Methodology

This section presents the metadata-driven heuristic method used to address the Document-to-Person Linking task. This method utilises document metadata to identify and rank matching person entities from the VRTI KG. Algorithm 1 summarises the procedure.

5.1. Metadata-driven Heuristic Method

The metadata-driven heuristic method is a rule-based candidate generation and ranking strategy for historical person linking. The method employs multi-criteria weighted similarity heuristics across several metadata dimensions, including personal names, temporal information, and geographic references. Given metadata extracted from State Papers Ireland documents, the approach retrieves and scores candidate entities from the reference KG, producing a ranked list of plausible matches.

Algorithm 1: Metadata-driven Heuristic Framework for Document-to-Person Linking

Input: Document D with metadata $\{name_d, date_d, L_d\}$; KG
Output: Ranked list of candidates C with composite scores
$C \leftarrow \emptyset$;
Normalize $name_d$ and document locations L_d;
foreach *entity* $e \in KG$ **do**
    /* 1. Name Similarity */
    $S_{name} \leftarrow \text{TokenSortRatio}(name_d, e.name)/100$;
    if $S_{name} < 0.5$ **then**
        continue ; // Discard weak lexical matches
    end
    /* 2. Temporal Consistency */
    if $date_d$ *and* $e.floruit$ *are available* **then**
        $S_{temp} \leftarrow (e.f_{lower} \leq date_d \leq e.f_{upper})?1.0 : 0.0$;
    end
    else
        $S_{temp} \leftarrow 0.5$;
    end
    /* 3. Location Similarity */
    if L_d *and* $e.locations$ *are available* **then**
        $S_{loc} \leftarrow \max\{\text{PartialRatio}(l_d, l_e) \mid l_d \in L_d, l_e \in e.locations\}/100$;
    end
    else
        $S_{loc} \leftarrow 0.5$;
    end
    /* 4. Composite Scoring */
    $Score_e \leftarrow (0.5 \cdot S_{name}) + (0.3 \cdot S_{temp}) + (0.2 \cdot S_{loc})$;
    Add $\{e, Score_e\}$ to C;
end
return *Sort* C *by* $Score_e$ *in descending order*;

This approach is specifically designed to address recurrent ambiguity patterns in historical archives. First, it accommodates orthographic variation and alternative name forms by applying fuzzy string similarity rather than exact lexical matching. This enables the system to link variant spellings or editorial forms of a person's name to a single canonical KG entity without requiring exhaustive alias enumeration. Second, it mitigates ambiguity between distinct individuals who share identical or highly similar names—such as family members active within overlapping time periods—by incorporating temporal compatibility scoring. By verifying whether the document date falls within a candidate's recorded floruit or lifespan interval, the method filters and ranks candidates according to historical plausibility. Third, it incorporates geographic similarity while assigning it a lower weight, thereby preventing document event locations (e.g., "Issued at" or "Sent to") from disproportionately penalising correct candidates whose recorded KG locations reflect biographical life events (e.g., birth or death place).

Unlike purely query-based retrieval mechanisms such as SPARQL exact matching, which depend on strict string equality and return unordered result sets, the heuristic method supports approximate matching, contextual plausibility assessment, and transparent relevance ranking. This combination allows the method to resolve ambiguous senders, overlapping identities, and metadata mismatches in a controlled and interpretable manner, making it particularly well-suited to historical archives where metadata is incomplete, inconsistent, or semantically heterogeneous.

Name Similarity Scoring

Name similarity functions as the primary filtering and scoring component, since personal names constitute the most immediate and discriminative identifier in documentary metadata. Both the document name and candidate entity names are preprocessed through normalisation (e.g., lowercasing and trimming of whitespace) to reduce superficial variation. Fuzzy string matching is then applied using a token-based similarity metric (token_sort_ratio), which accommodates reordered tokens and minor spelling differences. The similarity score is normalised to the range [0,1]:

$$name\_score = \frac{\text{TokenSortRatio}(doc\_name, entity\_name)}{100} \tag{1}$$

Candidates with a name similarity below 0.5 are discarded early to reduce computational cost and eliminate weak matches.

Temporal Consistence Scoring

Temporal compatibility is assessed by comparing the document creation date with the candidate entity's recorded floruit interval (floruit_lower, floruit_upper), which represents the period during which the individual was active.

If the document date falls within the floruit range, the temporal score is set to 1.0. If it falls outside, the score is set to 0.0. If temporal information is missing, a neutral default score of 0.5 is assigned to avoid introducing bias.

$$temporal\_score = \begin{cases} 1.0 & \text{if } floruit\_lower \leq doc\_date \leq floruit\_upper \\ 0.0 & \text{otherwise} \\ 0.5 & \text{if metadata unavailable} \end{cases} \tag{2}$$

This mechanism enforces temporal plausibility while remaining robust to missing data, and it helps disambiguate individuals who share the same name but lived or were active during different periods.

Location Similarity Scoring

Geographic similarity is computed between document-associated locations (issued_at, sent_to) and the candidate's biographical locations (birth and death locations) recorded in the KG. All locations are normalised to reduce superficial differences, and pairwise fuzzy partial matching is applied to account for variations in spelling or granularity. The maximum similarity across all document–entity location pairs is retained as the location score. When location metadata is unavailable, a neutral score of 0.5 is assigned. This approach allows partial matches, such as matching a city with its region, and treats geographic information as supporting context to help distinguish candidates, without strictly eliminating matches when document and KG locations do not exactly align.

$$location\_score = \begin{cases} \max\limits_{d \in \text{Doc location}, k \in \text{Entity location}} \dfrac{\text{PartialRatio}(d,k)}{100}, & \text{if } d \text{ and } k \text{ are available} \\ 0.5, & \text{otherwise} \end{cases} \tag{3}$$

By doing so, it reduces errors when document event locations (e.g., where a letter was sent or issued) differ from biographical locations in the KG, improving ranking accuracy for candidates whose names and dates match but whose location metadata does not perfectly align.

Composite Scoring and Candidate Ranking

The final candidate score is computed as a weighted linear combination of the three similarity components:

$$final\_score = 0.5 \cdot name\_score + 0.3 \cdot temporal\_score + 0.2 \cdot location\_score \tag{4}$$

These weights were selected heuristically based on domain knowledge of historical archives and the relative reliability of each metadata type. Names are generally the most precise, temporal information is moderately reliable, and locations are the least reliable due to semantic differences between document events and life events. Accordingly, higher weight is assigned to names (0.5), followed by temporal information (0.3), and then location (0.2). These values are used as initial heuristic settings and may be

refined through further empirical evaluation. While alternative weighting schemes could be applied, this combination balances precision and interpretability, producing robust candidate rankings even when metadata is incomplete or partially inconsistent.

6. Experiments and Results

This section evaluates the effectiveness of the metadata-driven heuristic method for Document-to-Person Linking. We compare its performance against a neural entity linking baseline to assess whether contextual embeddings provide advantages in this historical domain. The evaluation is conducted on the manually annotated State Papers Ireland documents using ranking-based metrics to measure disambiguation accuracy.

6.1. Dataset and Evaluation Setup

The evaluation was conducted on a corpus of 9,967 State Papers Ireland documents covering the period 1660–1715. The reference KG contains 175 candidate person entities representing individuals active during this period. 123 of them have multiple name variations.

To assess linking performance, a manually curated gold-standard dataset of 175 document–person linking instances was constructed. Each instance links a person's name appearing in document metadata fields (Sender and Recipient) to the corresponding KG URI. Only documents containing entities with verified ground-truth mappings were included in the evaluation.

The entity linking task is formulated as a ranking problem: given document metadata (name, date, and location), our method generates a ranked list of candidate KG entities. We report Top-1 accuracy (the highest-ranked candidate matches the ground-truth URI) and Top-3 accuracy (the ground-truth entity appears among the top three candidates).

6.2. Evaluation Metric

Performance is measured using Top-1 Accuracy, defined as:

$$Top-1\ Accuracy = \frac{\text{Number of correct top-ranked predictions}}{\text{Total evaluated instances}} \tag{5}$$

Similarly, Top-3 Accuracy is defined as:

$$Top{-}3\ Accuracy = \frac{\text{Number of instances where the ground-truth entity is among the top three candidates}}{\text{Total evaluated instances}} \tag{6}$$

For each evaluated document, both the Sender and Recipient fields are considered independently. A prediction is counted as correct when the URI of the highest-ranked candidate matches the manually annotated ground-truth URI.

This metric directly evaluates the effectiveness of the ranking strategy in selecting the correct entity without human intervention.

6.3. Results

To evaluate the metadata-driven heuristic method, we compared it against a BERT-based [22] neural entity linking baseline. Relevant document metadata (name, location and temporal) and KG entity properties (all name variants, associated places, and active years) were encoded using the bert-base-uncased model. Candidate entities were ranked based on the cosine similarity between document metadata and entity embeddings. For each document, similarity scores were calculated against all 175 candidate entities, and the top five candidates were retained. Performance was measured using Top-1 and Top-3 accuracy on the manually annotated ground-truth set.

Table 2 presents the evaluation results. The metadata-driven heuristic method achieved a Top-1 accuracy of 0.701 and Top-3 accuracy of 0.792, substantially outperforming the BERT-based baseline, which achieved 0.085 and 0.120, respectively. These results indicate that simple, interpretable heuristics are more effective than neural embeddings in this historical archival setting.

Method	Accuracy@1	Accuracy@3
Metadata-driven heuristic	0.701	0.792
Neural embedding-based(BERT)	0.085	0.120

Table 2
Top-1 and Top-3 accuracy of the metadata-driven heuristic versus a neural embedding-based (BERT) baseline for the Document-to-Person Linking task

6.4. Ablation Study

To assess the contribution of each metadata dimension to the overall heuristic method performance, we conducted an ablation study. We tested four variants:

- Full model: Uses name, temporal and location similarity combined according to the heuristic weighting scheme.
- No Temporal: Excludes temporal information from the composite score.
- No Location: Excludes location similarity from the composite score.
- Name only: Considers only name similarity for candidate ranking.

The results are summarised in the Table 3.

Model Variant	Accuracy@1
Full model	0.701
No Temporal	0.701
No Location	0.832
Name only	0.711

Table 3
Top-1 accuracy for different heuristic variants, illustrating the contributions of name, temporal, and location metadata to entity linking performance

The ablation results indicate that name similarity is the most discriminative attribute for linking in the State Papers Ireland corpus, while temporal information contributes little. Location similarity can be noisy or misleading, likely due to mismatches between document-event and person entity biographical KG locations (birth/death places). These findings justify the design of prioritising name similarity and assigning a lower weight to location information.

7. Discussion

The experimental results indicate that the BERT-based neural entity linking method did not outperform the metadata-driven heuristic method. While contextual embeddings have demonstrated strong performance in contemporary entity linking tasks, their effectiveness in this historical archival setting was limited.

Several factors may explain this outcome:

First, the metadata consists largely of structured archival fields (e.g., sender, recipient, issued location, and date) rather than rich narrative text. Transformer-based models such as BERT rely on contextual semantics learned from modern corpora and perform best when sufficient descriptive context is available. In this case, document representations were relatively short and formulaic, reducing the textual semantics available for embedding-based similarity.

Second, a key limitation of the neural entity linking method arises from the semantic mismatch between document metadata and KG entities. Document metadata (e.g., “Issued at,” “Sent to”, ”Date of Creation”) describes the document itself, whereas the KG entities’ properties describe the person (e.g., birth/death place, birth/death date). This difference in semantic scope means that embedding similarity between a document and a person entity often measures surface lexical overlap rather than meaningful entity correspondence. Heuristic method, by contrast, explicitly assign higher weight to the most discriminative properties (e.g., name similarity) and lower weight to noisier properties like location. BERT embeddings primarily capture co-occurrence patterns of words rather than the structural and role-based relations needed to disambiguate historical persons. As a result, embedding similarity alone fails to reliably link documents to the correct KG entity, especially when metadata fields do not overlap lexically with entity properties.

A promising direction for future research is the integration of NER. Rather than relying solely on name, temporal and location overlap, the system could apply NER to extract a person’s ORG, gender, family relation etc. from the document. Enriching KG entities with structured role information from Wikidata or biographical descriptions (e.g., offices, political roles, relations, working period) could also be fruitful.

Such structured enrichment could potentially improve disambiguation in cases where name and temporal information alone are insufficient. This approach may be particularly useful for challenging scenarios, such as ambiguous senders, father-son overlaps, and discrepancies between document event locations and KG biographical locations.

8. Conclusion and Future Work

This paper investigated the problem of linking historical documents from the State Papers Ireland (1660-1715) corpus to structured person entities in the VRTI KG using archival document metadata. The task is complicated by several characteristics of historical archives, including spelling variation in personal names, overlapping lifespan among individuals with identical names, and inconsistencies between document metadata and biographical information stored in KGs.

To address this challenge, we proposed a metadata-driven heuristic method that combines fuzzy name matching with temporal and location constraints to rank candidate entities. Unlike neural methods, our approach explicitly encodes domain knowledge through feature weighting, enabling greater transparency and controllability. We compared this approach with a BERT-based neural entity linking baseline. Evaluation shows that the heuristic method outperforms the neural model. The results suggest that embedding-based methods are not well-suited to this historical entity linking task. Importantly, this limitation is not due to the amount of text available but to the semantic mismatch between document metadata and KG entities. Document metadata (e.g., “Issued at,” “Sent to,” “Date of Creation”) describes the document itself, whereas KG properties describe the person (e.g., birth/death place, birth/death date).

Several limitations remain in this study. The method relies heavily on name similarity, which may not be sufficient in cases where multiple individuals share identical names and overlapping time periods. More generally, the approach depends on the semantic consistency of the metadata: features such as geographic information or event dates in archival documents can be noisy or inconsistent with KG biographical properties, which may reduce disambiguation performance.

In addition, the method is primarily designed and evaluated on the State Papers Ireland (1660–1715) corpus, and its generalisability to other historical datasets has not yet been systematically assessed. The weighting scheme is heuristic and was not exhaustively optimised or compared against a wide range of alternative strategies. Finally, while the approach is evaluated against a neural entity linking baseline, future work should include comparisons with additional heuristic and rule-based methods to provide a broader contextual evaluation.

Future work will focus on the integration of NER and information extraction techniques, identifying additional properties. We aim to enrich the linking process with structured evidence that more accurately

reflects the roles and activities of historical individuals.

Acknowledgments

This research was conducted with the financial support of Research Ireland under the ADAPT Centre for AI-driven Digital Content Technology which is funded under the Research Ireland Research Centres Programme (Grant 13/RC/2106_P2).

Declaration on Generative AI

During the preparation of this work, the author(s) used Grammarly to check grammar and spelling.

References

[1] E. Linhares Pontes, L. A. Cabrera-Diego, J. G. Moreno, E. Boros, A. Hamdi, N. Sidère, M. Coustaty, A. Doucet, Entity linking for historical documents: Challenges and solutions, in: International Conference on Asian Digital Libraries, Springer, 2020, pp. 215–231.

[2] A. Sabharwal, Digital curation in the digital humanities: Preserving and promoting archival and special collections, Chandos Publishing, 2015.

[3] M. A. M. Nandgude, Contemporary challenges and innovations in humanities research methodology, Contemporary Trends in Research (2025) 85.

[4] B. Yaman, A. Randles, L. McKenna, L. Kilgallon, D. Rincón-Yáñez, N. Johnston, P. Crooks, D. O'Sullivan, Expanding the virtual record treasury of ireland knowledge graph, Semant. Web J. (2024).

[5] A. Randles, L. McKenna, L. Kilgallon, B. Yaman, P. Crooks, D. O'Sullivan, The knowledge graph explorer for the virtual record treasury of ireland., in: VOILA@ ISWC, 2024, pp. 47–61.

[6] B. Yaman, L. McKenna, A. Randles, L. Kilgallon, P. Crooks, D. O'Sullivan, Digital prosopographical information in the virtual record treasury of ireland's knowledge graph for irish history, in: Proceedings of the 1st International Workshop of Semantic Digital Humanities (SemDH) Co-Located with the 21st Extended Semantic Web Conference, 2024.

[7] C. CIDOC, Conceptual reference model, URL: https://cidoc-crm. org (2003).

[8] H. Bast, N. Schnelle, Efficient and convenient sparql+ text search: A quick survey, in: Reasoning Web International Summer School, Springer, 2018, pp. 26–34.

[9] W. Shen, J. Wang, J. Han, Entity linking with a knowledge base: Issues, techniques, and solutions, IEEE Transactions on Knowledge and Data Engineering 27 (2014) 443–460.

[10] O.-E. Ganea, T. Hofmann, Deep joint entity disambiguation with local neural attention, in: Proceedings of the 2017 conference on empirical methods in natural language processing, 2017, pp. 2619–2629.

[11] L. Wu, F. Petroni, M. Josifoski, S. Riedel, L. Zettlemoyer, Scalable zero-shot entity linking with dense entity retrieval, in: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP), 2020, pp. 6397–6407.

[12] N. De Cao, G. Izacard, S. Riedel, F. Petroni, Autoregressive entity retrieval, arXiv preprint arXiv:2010.00904 (2020).

[13] M. P. K. Ravi, K. Singh, I. O. Mulang, S. Shekarpour, J. Hoffart, J. Lehmann, Cholan: A modular approach for neural entity linking on wikipedia and wikidata, in: Proceedings of the 16th conference of the european chapter of the association for computational linguistics: Main volume, 2021, pp. 504–514.

[14] A. Xin, Y. Qi, Z. Yao, F. Zhu, K. Zeng, B. Xu, L. Hou, J. Li, Llmael: Large language models are good context augmenters for entity linking, in: Proceedings of the 34th ACM International Conference on Information and Knowledge Management, 2025, pp. 3550–3559.

[15] Y. Benkhedda, A. Skapars, V. Schlegel, G. Nenadic, R. T. Batista-Navarro, Enriching the metadata of community-generated digital content through entity linking: an evaluative comparison of

state-of-the-art models, in: Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024), 2024, pp. 213–220.

[16] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al., Retrieval-augmented generation for knowledge-intensive nlp tasks, Advances in neural information processing systems 33 (2020) 9459–9474.

[17] M. Li, R. Qin, Dualgraphrag: A dual-view graph-enhanced retrieval-augmented generation framework for reliable and efficient question answering, Applied Sciences (2026).

[18] R. P. Gore, K. K. Ghosh, C. Sur, Ragner: Improving performance of llms using rag and specialized domain specific entity recognition, in: 2025 IEEE Guwahati Subsection Conference (GCON), IEEE, 2025, pp. 1–8.

[19] D. Nadeau, S. Sekine, A survey of named entity recognition and classification, Lingvisticae Investigationes 30 (2007) 3–26.

[20] M. Ehrmann, M. Romanello, S. Najem-Meyer, A. Doucet, S. Clematide, Overview of hipe-2022: named entity recognition and linking in multilingual historical documents, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2022, pp. 423–446.

[21] M. Piotrowski, Natural language processing for historical texts, Morgan & Claypool Publishers, 2012.

[22] J. Devlin, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).