Building Realistic Ground Truth Datasets of Personal Identification Information for Entity Matching

Ifeoluwapo Aribilola[1,3][0000−0003−4488−2925], Matteo Catena[2][0000−0002−5571−8269], Mamoona Asghar[1][0000−0001−7460−266X], John Breslin[3][0000−0001−5790−050X], and Renaud Delbru[2][0009−0007−7759−693X]

[1] School of Computer Science, College of Science and Engineering, University of Galway, University Road, Galway, H91 TK33, Ireland
ifeoluwapo.aribilola@universityofgalway.ie, mamoona.asghar@universityofgalway.ie
[2] Siren, 15 Market Street, Galway, H91 TCX3, Ireland
matteo.catena@siren.io, renaud.delbru@siren.io
[3] School of Engineering, College of Science and Engineering, University of Galway, University Road, Galway, H91 TK33, Ireland
john.breslin@universityofgalway.ie

Abstract. Entity matching (EM) is essential for connecting data across sources, particularly in sensitive domains like human trafficking investigations. However, research faces a critical gap: the lack of realistic gold standard datasets containing personal identifying information. This paper introduces a methodology for creating gold standard datasets, demonstrated through the development of a representative dataset for personal identification information (PII). Our approach combines multiple EM techniques to identify candidate matches, followed by a systematic annotation and validation process. Notably, our findings demonstrate that different techniques identify largely non-overlapping sets of matches, validating the need for our multi-technique methodology. Our approach provides a reproducible template for creating gold standard datasets in domains where realistic evaluation resources are scarce.

Keywords: Entity matching (EM) · Information Retrieval (IR) · Data Blocking · Data annotation · Human Trafficking · Large Language Models · Intelligence Investigations.

1 Introduction

Entity matching (EM) enables the identification and deduplication of records corresponding to the same real-world entities across various data sources [8]. Using approaches ranging from rule-based [15] to graph-based [5], probabilistic [23], and neural networks [1, 6], EM provides essential capabilities across domains including fraud detection [10] and risk assessment [24]. EM is particularly crucial in human trafficking investigations, where connecting disparate information can

reveal criminal networks operating across jurisdictions. Human trafficking – the unlawful exchange of individuals for exploitation through coercion, fraud, or force – involves "the use of force, threats, or other types of coercion, abduction, fraud, deception, or abuse of power" to exploit vulnerable individuals [26].

Background and Motivation: In investigative intelligence, particularly for human trafficking cases, EM faces unique challenges. Critical information is frequently dispersed among numerous unstructured or semi-structured sources. Perpetrators intentionally conceal identities using burner phones, aliases, fraudulent documents, and fictitious residences [12]. Accurate matching of these divergent pieces of information is crucial to identify key individuals or networks. Despite EM's critical importance in these sensitive domains, there is a severe lack of realistic, publicly accessible datasets containing personal identifying information (PII). The absence of realistic gold standard datasets with ground-truth labels severely constrains progress in developing robust EM solutions for real-world problems.

Problem Statement: Most benchmark datasets used in the EM literature are synthetic, highly curated, or semi-structured commercial product databases (e.g. Walmart-Amazon, DBLP-Scholar) [3, 16]. These datasets fail to capture the ambiguity and variability found in actual personal records, such as misspellings, nicknames, and inconsistent formatting. They rarely include realistic PII and often have artificially balanced label distributions. Furthermore, they fail to represent the complications found in trafficking investigations, where data is characteristically skewed, noisy, and incomplete.

The research community continues to focus on data situations that are excessively "clean" and domain-specific, restricting the generalisation of solutions, as pointed out by [18] in their assessment of blocking strategies. This disconnect between benchmark datasets and real-world applications creates a significant gap in our ability to develop and evaluate EM systems for high-stakes scenarios. Our research addresses these fundamental challenges through the following research questions:

RQ1: How do different EM approaches perform on PII?
RQ2: How can diverse entity matching techniques contribute to creating comprehensive gold standard datasets?

Contribution: In response to these research questions, this paper introduces a methodology for creating gold standard datasets, demonstrated through the development of a real-world PII dataset. We document an approach to dataset creation, annotation, and validation that can be reproduced in other domains. Our empirical results provide evidence that multiple complementary techniques are necessary for comprehensive gold standard creation, as different approaches identify largely non-overlapping sets of matches. We validate LLM-assisted annotation as an efficient approach for dataset creation, reducing the burden on domain experts while maintaining high quality. Through this process, we develop

a new gold standard dataset that captures the complexity of real-world entity matching challenges in human trafficking investigations.

Paper Organisation: This paper's remaining sections are arranged as follows: The relevant literature on entity matching pipelines, blocking techniques, and dataset creation challenges is discussed in Section 2. The dataset creation methodology is detailed in Section 3. Section 4 describes the complementary matching techniques incorporated in our dataset creation pipeline. The experimental validation of our methodology is presented in Section 5, while Section 6 analyses the evaluation results and discusses limitations. The conclusion and directions for future work are described in Section 7.

2 Related Works

This section discusses entity matching (EM) workflows and the limitations of existing benchmark datasets, particularly the lack of realistic personal information.

2.1 Entity Matching Workflows

EM workflows typically consist of multiple phases to manage scalability, heterogeneity, and noise while maintaining accuracy [2, 7]. These phases include data pre-processing to normalise inconsistencies, indexing (blocking) to create manageable candidate record pairs using techniques like canopy clustering [14], paired comparison with similarity functions, classification of pairs as matches or non-matches [9], and evaluation of result quality [2].

Recent advances in EM have leveraged deep learning [16] and foundation models [28], replacing traditional approaches with neural architectures such as DITTO [13]. Although these models perform well on benchmark datasets, they typically exhibit dramatic performance drops when applied to datasets with different properties [28], struggling with inconsistent schemas and varying levels of granularity. Models trained using supervised learning often overfit to specific features of their training data, including vocabulary, token patterns, and schema structures [27], resulting in poor generalisation to new domains.

2.2 Lack of Realistic Personal Information in Current Datasets

Robust EM is critical in high-stakes domains such as fraud detection, healthcare, and anti-human trafficking investigations, where personally identifiable information plays a key role. The research community requires datasets that include realistic PII, reflect diversity, and capture real-world complexity.

While EM research has benefited from standardised benchmark datasets [3, 16, 20], most suffer from significant limitations in domain coverage, data realism, and scalability. Most datasets used in EM research are synthetic or excessively

cleaned, failing to represent the complexity, diversity, and noise present in operational situations [18]. Authentic personal information typically exhibits inconsistencies in formatting, incomplete fields, and cultural/linguistic variations – characteristics essential for evaluating EM systems but largely absent from existing benchmarks.

Few publicly available datasets contain authentic personal information suitable for EM research. Sources like the ICIJ Offshore Leaks Database and public corporate registry data contain real names, addresses, and organisational affiliations with authentic noise and inconsistencies. However, these datasets lack reliable ground truth for training and evaluation, highlighting the need for methodologies to transform such sources into usable gold standards.

In this paper, we introduce a new dataset constructed from two publicly available sources: (1) a subset of OpenCorporates[4] focused on entities of type person (company officers) and (2) a subset of the ICIJ Offshore Leaks Database [11], also filtered on personal entities. Our methodology emphasises the preservation of real-world variability in names and addresses, offering a more representative testbed for EM tasks than existing synthetic datasets. This approach provides a template for generating additional datasets for various EM scenarios beyond our specific implementation.

3 Dataset Creation Methodology

This section presents our methodology for creating gold standard datasets for entity matching (EM), demonstrated through the development of a dataset involving personal information. While our implementation focuses on specific sources, the methodology is generalisable to other domains. As illustrated in Fig. 1, our approach employs multiple complementary techniques to ensure comprehensive coverage of potential matches. Instead of relying on a single matching method – which could systematically miss certain types of matches – we leverage three distinct approaches to generate candidate pairs, which then undergo expert validation to create a gold standard dataset that captures a larger spectrum of matching challenges.

3.1 Data Sources

The ICIJ Offshore Leaks Database, released in 2013, contains information on offshore entities and shell corporations derived from document leaks, including the Panama Papers and Paradise Papers. The dataset includes information on officers, entities, and addresses in a semi-structured format with real-life discrepancies and multiple languages. The second source is the officers' dataset from the OpenCorporates platform, containing publicly available information on officials and directors connected to corporations, including names, positions, and partial addresses.

[4] `https://opencorporates.com`

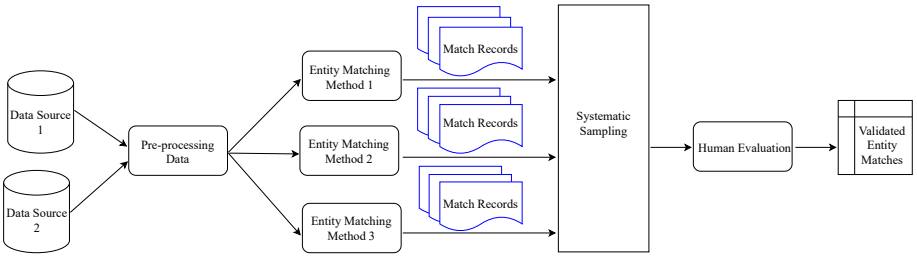


Fig. 1: The process flow for the dataset creation, showing data sources feeding into multiple entity matching methods, followed by systematic sampling and human evaluation to produce validated entity matches.

These sources are particularly valuable for our methodology because they a) contain authentic personal identifiers (names, addresses) with real-world variations, b) include the natural noise and inconsistencies found in actual records, c) represent scenarios relevant to investigative domains, and d) are publicly available, allowing for reproducible research.

For this experiment, we extracted only person entities, totaling 567,707 from ICIJ and 4,689,708 from OpenCorporates.

3.2 Data Pre-processing

An initial step is transforming raw data into formats suitable for EM while preserving the authentic variability that makes the dataset valuable for evaluation.

Selective Attribute Extraction: We first identify and extract the most relevant attributes for matching. Each dataset contains numerous attributes that are not essential for the matching task; keeping these would increase computational requirements and potentially introduce noise. For our implementation, we selected a) fields `Name` and `Address` from the OpenCorporates dataset, and b) fields `Name` and `Addresses.postalAddress` from the ICIJ dataset.

Structure Normalisation: The entities in both datasets were stored in JSON format with nested structures, requiring normalisation to facilitate processing. We flatten the JSON structure by i) converting nested objects into dot notation keys (e.g., `Address.Street`), and ii) aggregating arrays into lists (e.g., `''Emails.Type'': ['Work'', ''Home'']`). Listing 1.1 shows an example JSON entity before normalisation, and Listing 1.2 shows the same entity after normalisation. This transformation standardises the structure while preserving all information.

3.3 Multi-Technique Candidate Generation

A core contribution is using multiple complementary techniques to generate candidate pairs. This approach addresses a fundamental limitation of single-

Listing 1.1: Before JSON normalization

```
{
  ''Name'': ''John Doe'',
  ''Address'': {''Street'': ''123 Main St'', ''City'': ''Metropolis''},
  ''Emails'': [
    {''Type'': ''Work'', ''EmailAddress'': ''john.work@example.com''},
    {''Type'': ''Home'', ''EmailAddress'': ''john.home@example.com''}
  ]
}
```

Listing 1.2: After JSON normalization

```
{
  ''Name'': ''John Doe'',
  ''Address.Street'': ''123 Main St'',
  ''Address.City'': ''Metropolis'',
  ''Emails.Type'': [''Work'', ''Home''],
  ''Emails.EmailAddress'': [''john.work@example.com'', ''john.home@example.
      com'']
}
```

technique methods: each has inherent biases that can systematically miss certain types of matches. Our implementation employs three distinct techniques (detailed in Section 4):

- **Rule-based blocking**: Using the RecordLinkage library to identify candidates based on explicit similarity thresholds for names and addresses
- **Information retrieval with LLM matching**: Combining Lucene-based blocking with large language model assessment to identify semantically similar entities
- **In-house dataset**: Incorporating previously identified potential matches

Each technique provided candidate pairs labeled as potential matches (`Yes`), potential non-matches (`No`), or uncertain cases (`Maybe`), which are merged for expert validation. The value of this multi-technique approach is that each method identifies largely non-overlapping sets of matches (see Section 5).

3.4 Sampling Strategy for Expert Validation

Following the EM process (see Section 4), three techniques were applied to generate matched record pairs: Lucene+LLM (LL), the in-house dataset (IH), and RecordLinkage (RL). The LL pipeline produced 19,637 matched records, IH included 4,199, and RL produced 170 matches, collectively generating 24,006 matched records before deduplication.

After removing duplicates, we had 21,963 unique matched records. To create a manageable dataset for annotation while ensuring representation of all sources, we implemented systematic sampling that: i) includes all matches from the rule-based blocking (RL) due to their small number, ii) selects a balanced proportion

from the remaining sources (LL and IH), and iii) ensures all three sources are represented proportionally in the final subset. This systematic approach produced a balanced dataset of 500 candidate pairs that preserves each technique's unique contribution and reflects diverse matching behaviors across methods.

3.5 Expert Validation and Gold Standard Creation

The final phase of our methodology is the expert validation process, which transforms candidate pairs into a reliable gold standard. The experts compare the matched entities retrieved from the samples and decide if they truly match (`Yes`), are different (`No`), or are indecisive (`Maybe`).

Annotation Platform Setup: We selected Label Studio [25] for its flexibility in supporting manual annotation. The platform was configured to a) display paired entity records with relevant attributes, b) allow classification as matches, non-matches, or uncertain, c) randomise assignment to different annotators, and d) track decisions and identify conflicts.

Annotator Recruitment and Guidelines: Six domain experts with experience in investigative intelligence were recruited. Each received guidelines explaining a) the definition of a match in human trafficking investigations; b) how to assess name variations; c) how to evaluate address discrepancies; and d) when to mark cases as uncertain.

Conflict Resolution: A key element is the systematic resolution of annotation conflicts. When annotators disagreed on classifications, we i) applied Krippendorff's alpha [17] to measure agreement, flagging instances with zero agreement; ii) had a panel of three experts jointly review each conflict; iii) performed additional verification when necessary, including using Google Maps to confirm address equivalence; and iv) reached consensus through discussion and additional evidence.

Gold Standard Compilation: The final gold standard dataset combined a) pairs with unanimous annotator agreement, and b) pairs with resolved conflicts through the expert panel process. This gave a total of 500 candidate pairs, with 474 pairs as `Yes`, 9 pairs as `No`, and 17 pairs as `Maybe`.

4 Matching Methodology

This section describes the complementary techniques used in our dataset creation pipeline to generate candidate pairs. Each technique contributes unique matches to ensure comprehensive coverage of entity relationships, addressing how diverse approaches can create comprehensive gold standard datasets (RQ2).

4.1 Rule-Based Approach with RecordLinkage

The first component employs rule-based matching using the Python RecordLinkage Toolkit [4], with sequential indexing, comparison, and classification steps.

Blocking To manage computational complexity, we applied the RecordLinkage library's blocking functionality on the `Name` and `Address` fields from the OpenCorporates dataset, matched with the `Name` and `Addresses.postalAddress` fields from the ICIJ dataset. This process generated 35,195 candidate pairs, significantly reducing the comparison space from the original data sources.

Comparison Each candidate pair was compared using string similarity metrics to generate a comparison vector. We applied the Jaro-Winkler similarity metric (threshold 0.85) to names due to its effectiveness with character matches, transpositions, and common prefixes. For addresses, we used the Damerau-Levenshtein distance (threshold 0.7) to accommodate typographical errors commonly found in address data. We also conducted alternative comparisons using Jaro-Winkler on both fields. Notably, these specific string comparison functions were critical – without them, the algorithm returned no matches.

Classification The candidate pairs were filtered to retain only those in which the sum of similarity scores exceeded 1.0, that is, at least two fields exhibited similarity above the defined thresholds. This simple rule-based classification produced a total of 170 matches.

This approach excels at identifying exact or near-exact matches with consistent formatting but may miss semantically equivalent matches with significant syntactic differences.

4.2 IR-Based blocking with LLM-Based Matching

The second approach combines information retrieval with the assessment of large language models, leveraging both lexical similarity and semantic understanding. As shown in Fig. 2, this process consists of two main stages: 1) an N-Gram Blocking phase using IR techniques, and 2) an LLM-based Matching phase.

N-Gram Blocking with Lucene Inspired by Sparkly [19], we use Lucene to index the OpenCorporates dataset with an n-gram analyser (3-gram tokens) to capture partial matches despite spelling or formatting variations. Each entity in the smaller dataset (ICIJ dataset) was then used to probe against the OpenCorporates index. The process included:

- **Blocking Parameters:** We performed blocking on the name and address fields from both datasets.
- **Document Representation:** We concatenated attribute values into a single text string for each entity, creating a "bag of n-grams" without field boundaries.

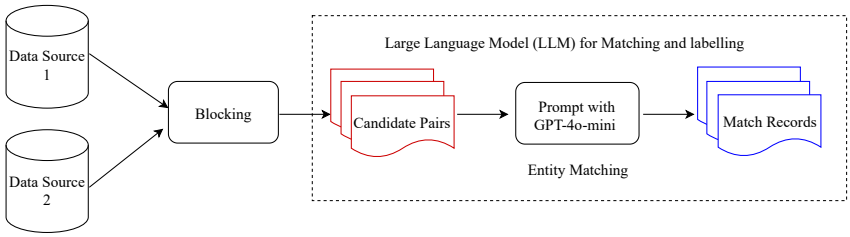


Fig. 2: IR-based blocking with LLM-based matching workflow: Lucene identifies initial candidates based on lexical similarity, then GPT-4o-mini assesses semantic equivalence.

- **Query Construction:** For each ICIJ entity, we similarly created a query string as a disjunction of terms.
- **Candidate Ranking:** We used BM25 [22] scoring to rank matches, retrieving the top-5 candidates for each query.

LLM-based Matching The second stage used GPT-4o mini [21] to assess candidate pairs' semantic similarity. We used OpenAI's Batch API[5] with zero temperature and fixed seed for deterministic outputs. Each pair was evaluated using the prompt in Listing 1.3, classifying them as matches (`Yes`), non-matches (`No`), or uncertain (`Maybe`).

Listing 1.3: Zero-shot prompt for entity matching with GPT-4o-mini

```
Compare the following two entity descriptions and determine if they refer
    to the same real-world entity.
  Entity 1: '{entity_1}'
  Entity 2: '{entity_2}'
Decide as follows:
  ''Yes'' if you are almost certain they refer to the same entity.
  ''Maybe'' if there is uncertainty.
  ''No'' if you are certain they refer to different entities.
Respond with only one token: ''Yes'', ''Maybe'', or ''No'', with no
    additional text.
```

This approach offers unique advantages: It can identify semantically equivalent matches despite syntactic differences (e.g., "J. Smith" vs. "Jonathan Smith, Esq.") and requires no manually crafted rules or labeled training data. Initial analysis of `Yes`-labeled pairs showed promising accuracy. The approach is economically feasible, with an estimated cost of approximately $57 for processing 3 million candidate pairs, and is highly scalable.

[5] `https://platform.openai.com/docs/api-reference/batch`

4.3 In-house Dataset

Our methodology incorporated a third source of candidate pairs from an existing repository of potential matches between data sources. This repository contained previously identified potential entity matches, contributing significantly to our candidate generation process. The inclusion of this established collection provided complementary candidate pairs that might be missed by algorithmic approaches, particularly those involving complex variations in personal identifiers that often require domain expertise to recognize.

4.4 Complementary Value of Multiple Techniques

Each approach brings distinct strengths to our dataset creation methodology. *Rule-based matching* excels at identifying exact or near-exact matches with consistent formatting. *IR & LLM matching* identifies semantically equivalent entities despite syntactic differences, operating effectively in a zero-shot setting. The *in-house dataset* provides additional matching candidates that complement the algorithmic approaches.

Our analysis in Section 5 confirms that these approaches identify largely non-overlapping sets of matches, validating our multi-technique methodology as essential for creating comprehensive gold standard datasets. By combining these complementary approaches and subjecting their outputs to expert validation, we create a gold standard that captures a broader spectrum of matching patterns than would be possible with any single technique.

5 Dataset Evaluation

This section evaluates our dataset creation methodology, focusing on how effectively our multi-technique approach produces a comprehensive gold standard. Rather than competitively comparing techniques, we analyse how different approaches contribute complementary matches to the dataset, addressing our second research question (RQ2). In addition, the gold standard dataset helped us accurately identify the strength of the entity matching techniques, thus answering our RQ1.

5.1 Experimental Setup

All experiments were conducted within a controlled software environment on a Windows 10 operating system, using Python 3.10 with Pandas 2.1.4, RecordLinkage 0.16, and Lucene 9.6.

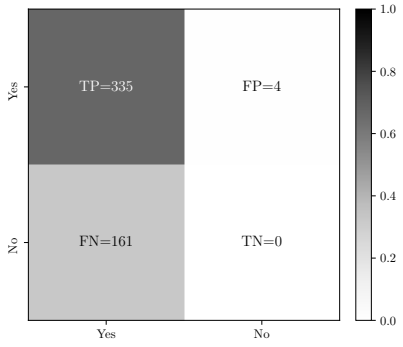
5.2 Contribution Analysis of Multiple Techniques

The core hypothesis of our methodology is that different techniques identify largely non-overlapping sets of matches. To evaluate this hypothesis, we analysed

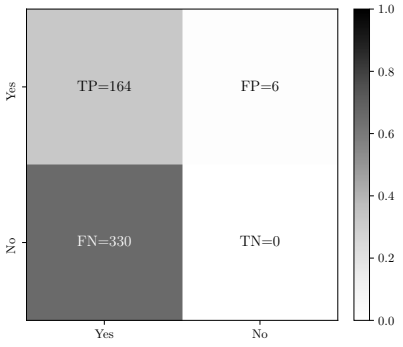
the overlap in matches identified by our three approaches: rule-based RecordLinkage (RL), the in-house dataset (IH), and Lucene with LLM matching (LL). Table 1 presents the distribution of matches identified by each technique individually and in combination, while Fig. 3 shows the confusion matrix for these distributions. The true positive (TP) is the number of actual `Yes` cases detected by the EM techniques; the false positive (FP) is the number of records these techniques detected as `Yes` but they are actually `No` or `Maybe` cases, while the false negative (FN) is the records that a technique did not detect but the other techniques detected.

Table 1: Distribution of matches identified as `Yes` cases by each entity matching technique individually and in combination.

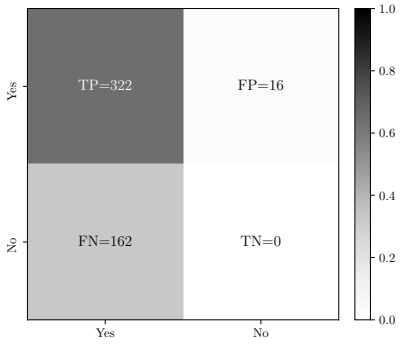
RL only	IH only	LL only	RL&IH	RL&LL	LL&IH	RL&IH&LL
20	110	110	32	31	110	87



(a) IH



(b) RL



(c) LL

Fig. 3: Confusion matrix for entity matching techniques result.

Only 87 matches (17.4% of the total) were identified by all three approaches, demonstrating that no single technique could have created a comprehensive gold standard on its own. Each approach contributed unique matches: RL exclusively identified 20 matches (4%), IH contributed 110 matches (22%), and LL found 110 matches (22%). The significant overlap between IH and LL (110 matches) indicates that zero-shot LLM matching can identify many of the same patterns as domain-specific sources. However, both approaches missed matches captured by rule-based methods, confirming that different techniques have complementary strengths.

5.3 Performance Characteristics of Dataset Creation Components

To understand the quality of contribution of each approach, we evaluated their performance against the expert-validated gold standard. Table 2 presents the precision and recall metrics for each technique. All three approaches maintained high precision (>95%), but differed significantly in recall: RL achieved 33.20% recall, while IH reached 67.54% and LL 66.53% respectively. Even the best-performing technique missed approximately one-third of true matches, validating our methodology's focus on combining multiple high-precision approaches to maximise recall.

Table 2: Contribution characteristics of different techniques to gold standard creation. Each approach demonstrates distinct precision-recall trade-offs, highlighting the value of a multi-technique methodology.

	RL	IH	LL
Precision	96.47%	98.82%	95.27%
Recall	33.20%	67.54%	66.53%

6 Discussion

This section synthesizes key insights from our experimental results and discusses our findings, implications, and limitations.

Performance and Complementarity of Matching Techniques: Our evaluation reveals distinct strengths across entity matching techniques. The RL method achieves high precision (96.47%) but low recall (33.20%), reflecting its emphasis on accuracy over coverage. The IH collection delivers both high precision (98.82%) and better recall (67.54%), highlighting the value of domain-specific expertise. Notably, the LL model performs competitively with 95.27% precision and 66.53% recall, despite not requiring domain tuning. Importantly, only 17.4% of the matches was identified by all three methods, validating our core claim: no single approach captures all valid matches. This confirms the necessity

of a multi-technique methodology to construct comprehensive and realistic gold standard datasets.

Our evaluation reveals distinct strengths across entity matching techniques. The RL method achieves high precision (96.47%) but low recall (33.20%), reflecting its emphasis on accuracy over coverage. The IH approach delivers both high precision (98.82%) and better recall (67.54%). Notably, the LL model performs competitively with 95.27% precision and 66.53% recall, despite not requiring domain tuning. Importantly, only 17.4% of the matches was identified by all three methods, validating our core claim: no single approach captures all valid matches. This confirms the necessity of a multi-technique methodology to construct comprehensive gold standard datasets.

Impact of Blocking Parameters: Our findings suggest that blocking parameters may significantly impact matching outcomes. Using Lucene blocking with a top-k candidate (e.g., $k = 5$) enhances efficiency but imposes a recall ceiling, regardless of how advanced the matching algorithm is. Although increasing k could improve recall, it also increases computational costs in the matching phase. This underscores a central challenge in entity matching: balancing efficiency with the risk of missing true matches. In human trafficking investigations, where missing a match has serious consequences, this trade-off is especially critical. Future implementations might benefit from adaptive top-k selection based on similarity score distributions, as suggested in the Sparkly system and other recent approaches.

Effectiveness of LLM-based Matching: Our results demonstrate that LLMs can effectively serve as matching components in entity resolution pipelines, even without domain-specific fine-tuning. The LL approach achieves competitive precision and recall compared to other methods, despite operating in a zero-shot setting. This suggests that LLMs capture generalisable matching criteria that transfer well across domains. The relatively high precision achieved by the LLM approach (95.27%) indicates that foundation models effectively leverage pretrained knowledge about names, addresses, and entity relationships to make accurate matching decisions. LLMs can be valuable tools for preliminary annotation in dataset creation, reducing the manual effort required from domain experts.

Dataset Contributions and Limitations: The gold standard dataset developed through our methodology fills a critical gap in entity matching research by incorporating realistic personal identifying information. Using real-world open data sources (ICIJ and OpenCorporates), we applied multiple matching techniques with expert validation to create a dataset that reflects the complexity found in real-world scenarios. This demonstrates a practical and reproducible approach for building high-quality evaluation datasets. However, the dataset has limitations: it includes only 500 records, which is relatively small, and it reflects specific geographical and linguistic contexts.

Implications for Real-world Applications: Our findings have several methodological implications. First, the complementary nature of different matching approaches underscores the importance of employing multiple techniques when creating gold standard datasets to ensure comprehensive coverage. Second, our methodology provides a template for developing domain-specific gold standards that can be adapted to other investigative contexts. Third, the competitive performance of LLM matching demonstrates the potential of foundation models to accelerate dataset creation with minimal domain-specific tuning.

7 Conclusion and Future Work

Entity matching is a cornerstone of data integration, especially in high-stakes domains such as human trafficking investigations, where the accurate linkage of personally identifying information is critical. Despite the progress in algorithmic development, a significant bottleneck in advancing entity matching research lies in the scarcity of realistic gold standard datasets that involve personal data and reflect true-world complexity.

This paper addresses this gap by proposing and validating a reproducible methodology for constructing gold standard datasets that contain personal identifying information. By integrating multiple entity matching techniques, we effectively broadened the discovery of matching candidate pairs and demonstrated that these techniques often surface complementary, non-overlapping matches. This insight underscores the limitations of relying on a single matching approach and supports the use of ensemble methods in dataset curation.

Our contributions include: (1) the creation of a gold standard dataset creation that incorporates realistic personal identifying information (Section 3), (2) the application of multiple complementary entity matching techniques to ensure comprehensive coverage of potential matches (Section 4), and (3) empirical validation showing the added value of combining diverse strategies in capturing comprehensive match sets (Section 5).

Our results demonstrate that LLMs are promising tools for entity matching. Even in zero-shot settings, the LL model performed on par with domain-specific approaches – achieving 95.27% precision – by leveraging generalizable pre-trained knowledge of names, addresses, and relationships. LLMs can serve as effective annotation aids, helping reduce the burden on human experts during dataset creation while maintaining high quality.

In future work, we plan to apply this methodology to larger datasets, incorporating additional entity matching techniques to further improve coverage and representativeness. We also intend to explore how blocking parameters might affect overall recall in different contexts, which could help optimize the balance between computational efficiency and match discovery. The methodology could be extended to create gold standard datasets in other domains. Finally, we plan to leverage the constructed dataset to support the development of anti-human trafficking initiatives.

The created dataset is made available to the research community at the GitHub repository `https://github.com/rendel/pii_match`. If you use this dataset in your research or publication, please cite this paper/repository and acknowledge the original data sources.

Acknowledgments. This publication has emanated from research supported in part by the European Digital Innovation Hub Data2Sustain, co-funded by Ireland's National Recovery and Resilience Plan (the EU's Recovery and Resilience Facility), the Digital Europe Programme, and the Government of Ireland.

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Barlaug, N., Gulla, J.A.: Neural networks for entity matching: A survey. ACM Trans. Knowl. Discov. Data **15**(3) (2021)
2. Christen, P.: Data matching. Springer (2012)
3. Das, S., Doan, A., G. C., P.S., Gokhale, C., Konda, P., Govind, Y., Paulsen, D.: The magellan data repository. `https://sites.google.com/site/anhaidgroup/projects/data`
4. De Bruin, J.: Python record linkage toolkit: A toolkit for record linkage and duplicate detection in python (2019)
5. Devezas, J., Nunes, S.: A review of graph-based models for entity-oriented search. SN Comput. Sci. **2**(6) (2021)
6. Gaur, B., Saluja, G.S., Sivakumar, H.B., Singh, S.: Semi-supervised deep learning based named entity recognition model to parse education section of resumes. Neural Comput. Appl. **33**(11) (2021)
7. Getoor, L., Machanavajjhala, A.: Entity resolution: theory, practice & open challenges. Proc. VLDB Endow. **5**(12) (2012)
8. González-Gallardo, C.E., Tran, H.T.H., Hamdi, A., Doucet, A.: Leveraging open large language models for historical named entity recognition. In: Linking Theory and Practice of Digital Libraries (2024)
9. Gu, L., Baxter, R.: Decision Models for Record Linkage, pp. 146–160. Springer Berlin Heidelberg, Berlin, Heidelberg (2006). `https://doi.org/10.1007/11677437_12`, `https://doi.org/10.1007/11677437_12`
10. Gupta, A.: Detection of spam and fraudulent calls using natural language processing model. In: 2024 Sixth International Conference on Computational Intelligence and Communication Technologies (CCICT) (2024)
11. International Consortium Investigative Journalists: Icij offshore leaks database (2021), `https://offshoreleaks.icij.org/`
12. Li, Y., Nair, P., Pelrine, K., Rabbany, R.: Extracting person names from user generated text: Named-entity recognition for combating human trafficking. In: Findings of the Association for Computational Linguistics: ACL 2022 (2022)
13. Li, Y., Li, J., Suhara, Y., Doan, A., Tan, W.C.: Deep entity matching with pre-trained language models. Proc. VLDB Endow. **14**(1) (2020)
14. McCallum, A., Nigam, K., Ungar, L.H.: Efficient clustering of high-dimensional data sets with application to reference matching. In: Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2000)

15. Moore, S., Nguyen, H.A., Chen, T., Stamper, J.: Assessing the quality of multiple-choice questions using gpt-4 and rule-based methods. In: Responsive and Sustainable Educational Futures (2023)
16. Mudgal, S., Li, H., Rekatsinas, T., Doan, A., Park, Y., Krishnan, G., Deep, R., Arcaute, E., Raghavendra, V.: Deep learning for entity matching: A design space exploration. In: Proceedings of the 2018 International Conference on Management of Data (2018)
17. Nanayakkara, C., Christen, P., Christen, V.: Unsupervised evaluation of entity resolution. ACM J. Data Inf. Qual. **17**(1), 1–31 (Mar 2025)
18. Papadakis, G., Skoutas, D., Thanos, E., Palpanas, T.: Blocking and filtering techniques for entity resolution: A survey (2020)
19. Paulsen, D., Govind, Y., Doan, A.: Sparkly: A simple yet surprisingly strong tf/idf blocker for entity matching. Proc. VLDB Endow. **16**(6) (2023)
20. Primpeli, A., Peeters, R., Bizer, C.: The wdc training dataset and gold standard for large-scale product matching. In: Companion Proceedings of The 2019 World Wide Web Conference (2019)
21. Radford, A., Narasimhan, K.: Improving language understanding by generative pre-training. In: Computer Science, Linguistics (2018)
22. Robertson, S., Zaragoza, H.: The probabilistic relevance framework: Bm25 and beyond. Found. Trends Inf. Retr. **3**(4), 333–389 (Apr 2009). https://doi.org/10.1561/1500000019, https://doi.org/10.1561/1500000019
23. Ruz, G.A., Henríquez, P.A., Mascareño, A.: Bayesian constitutionalization: Twitter sentiment analysis of the chilean constitutional process through bayesian network classifiers. Mathematics **10**(2) (2022)
24. Shishehgarkhaneh, M.B., Moehler, R.C., Fang, Y., Hijazi, A.A., Aboutorab, H.: Transformer-based named entity recognition in construction supply chain risk management in australia. IEEE Access **12** (2024)
25. Tkachenko, M., Malyuk, M., Holmanyuk, A., Liubimov, N.: Label Studio: Data labeling software (2020-2025), https://github.com/HumanSignal/label-studio, open source software available from https://github.com/HumanSignal/label-studio
26. United Nations: Protocol to prevent, suppress and punish trafficking in persons especially women and children, supplementing the united nations convention against transnational organized crime. https://www.ohchr.org/en/instruments-mechanisms/instruments/protocol-prevent-suppress-and-punish-trafficking-persons (Nov 2000), accessed: 2025-4-17
27. Wu, R., Chaba, S., Sawlani, S., Chu, X., Thirumuruganathan, S.: Zeroer: Entity resolution using zero labeled examples. In: Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data (2020)
28. Zhang, Z., Groth, P., Calixto, I., Schelter, S.: A deep dive into cross-dataset entity matching with large and small language models. In: Proceedings of the 28th International Conference on Extending Database Technology (2025)