# A Hybrid Method for Standardising Civil Registration and Vital Statistics (CRVS) Location Data

**I Sandyawan[1], Y Rimawati[2*], and A Rismansyah[2]**

[1] The University of Melbourne, Australia
[2] BPS-Statistics, Indonesia

*Corresponding author's email: yeni.rima@bps.go.id

**Abstract.** Civil Registration and Vital Statistics (CRVS) systems in archipelagic contexts like Indonesia face persistent challenges in location data standardisation due to free-text entries that vary in spelling, formatting, and granularity. This study introduces a multi-stage hybrid framework that systematically converts these unstructured entries into official administrative codes using deterministic matching, fuzzy probabilistic matching, and geocoding. This study processed 841,126 birth and death records using Python (Pandas, RapidFuzz, Geopy). Cumulatively, all stages achieved a combined match rate of 85.44% for births and 67.12% for deaths. The layered pipeline ensured speed, precision, and coverage for real-world CRVS data. The findings demonstrate enhanced geographic precision in vital statistics, enabling more reliable public health and demographic applications. Future improvements may include transformer-based embeddings, active learning for ambiguous records, and uncertainty-aware geocoding techniques. This framework establishes a scalable, robust pathway for elevating the granularity and reliability of geolocated vital event data.
**Keywords:** CRVS standardisation, fuzzy record linkage, geocoding, spatial proximity analysis, vital statistics

## 1. Introduction

Accurate and reliable vital statistics are the bedrock of effective public health planning, demographic analysis, and evidence-based policymaking worldwide [1]. These statistics, capturing critical life events like births and deaths, offer indispensable insights into population dynamics, health trends, and societal needs [2][3][4]. However, the quality and usability of vital statistics are intrinsically tied to the precision and consistency of the underlying civil registration records. A pervasive challenge in national Civil Registration and Vital Statistics (CRVS) systems, particularly relying on open-ended input fields, is the inherent variability and ambiguity of location data. This issue is particularly acute in large, diverse archipelagic nations like Indonesia, where local nuances in placename recording frequently lead to inconsistencies [5].

In Indonesia, the standardisation of administrative regions is primarily governed by the Ministry of Home Affairs (MOHA), which issues official administrative codes for provinces (e.g., 32 for Jawa Barat), regencies/cities (e.g., 3201 for Kabupaten Bogor), districts (e.g., 320129 for Kecamatan

Ciomas), and villages/kelurahan (e.g., 3201292006 for Desa Pagelaran). While Statistics Indonesia (BPS) also maintains its own Statistical Working Areas for data collection and dissemination, theMOHA codes are fundamental for official administrative purposes and vital registration. It is important to note that a mapping exists between BPS codes and MOHA administrative codes, indicating an inter-agency need for data alignment and consistency.

Even though MOHA has a standardised code for administrative regions, some variables for vital event registration still collect unstructured place of event information[6] for simplicity reasons when getting data from citizens. This open-ended data collection method inevitably leads to inconsistencies in spelling, format, and nomenclature, creating a significant hurdle for data aggregation and analysis. For instance, the same location might be recorded in multiple ways, or a generic name could refer to several distinct places across different administrative regions. Such variability directly compromises the ability to accurately assign events to standardised administrative boundaries, diminishing the reliability of derived vital statistics for geographic-specific planning and reporting.

Traditional methods for standardising location data often fall short [7][8]. Purely deterministic approaches, while computationally efficient and effective for exact string matches, frequently fail to accommodate the variability commonly found in human-entered text. Such variability includes typographical errors, inconsistent formatting, and alternative spellings, which lead to lower match rates and overlooked valid records. Conversely, relying exclusively on external geocoding services, despite their advanced algorithms and access to extensive geographic databases, often results in ambiguous or imprecise matches. This ambiguity arises due to generic place names, overlapping geographic boundaries, and entries with insufficient or unclear information, thereby necessitating additional manual effort for disambiguation and verification. These inherent drawbacks in each method highlight the need for a hybrid or multi-stage approach that balances speed, accuracy, and robustness when standardising diverse real-world location data.

Recent frameworks have achieved impressive performance in related address-matching tasks. For example, the FLAP system (Framework for Linking free-text Addresses to the UPRN database) demonstrated an adjusted matching accuracy of 0.992 when linking unstructured UK addresses to UPRN records, showing robustness to typographical and formatting variants [6]. Deep learning–driven libraries such as DeezyMatch offer flexible, neural-based fuzzy matching and candidate ranking, particularly useful for multilingual or noisy datasets like CRVS entries [9]. Meanwhile, classical string similarity algorithms (e.g., Cosine, Dice, LCS) remain relevant, especially when enhanced using safety-focused classification layers to reduce false positives in sensitive applications [10]. Hybrid methodologies that combine textual similarity, geospatial features, and conditional dependencies have also shown improved match coverage in practical scenarios [11][12].

To bridge the methodological gap in CRVS location standardisation, this paper proposes a multi-stage hybrid framework that integrates deterministic, probabilistic matching, and geocoding. By calculating distances to administrative centroids, this framework aims to resolve ambiguity and enhance location assignment accuracy, transforming raw entries into high-quality, standardised data that bolsters the foundation of vital statistics for public sector applications.

## 2. Research Method

This section details the proposed multi-stage hybrid framework for standardising unstructured location data in civil registration systems (see figure 1). The methodology systematically transforms free text location entries into standardised administrative codes, enhancing vital statistics quality. This approach integrates deterministic matching, probabilistic-statistical matching, and geocoding using Python libraries such as Pandas, RapidFuzz, and Geopy [13], forming a robust pipeline for comprehensive data standardisation.

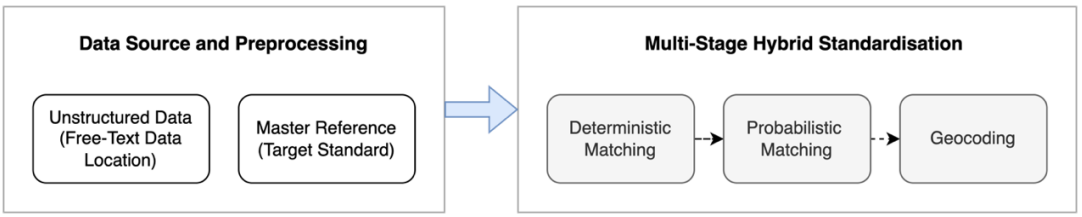


**Figure 57.** Pipeline Overview: From Free Text Data to Hybrid Standardisation

### *2.1. Data Source and Preprocessing*

This study uses two data sources: unstructured "place of event" entries from CRVS (births and deaths) and a standardised master list of administrative regions. The data used in this study were obtained from MOHA records from 2019 to 2024. This data is obtained in a CSV format and handled securely by CRVS teams without any personal identifiable information, with the goal of calculating vital statistics based on registration events. The "place of event" data for deaths and births are sourced from free-text inputs, making it difficult to standardise. In contrast, the master regional data is a standardised dataset used for matching with the vital event data.

Robust text normalisation starts with properly extracting references from unstructured place names. The geoparsing field recently consolidated diverse methods (rule-based, gazetteer matching, statistical, and hybrid) for pulling location references from free text, as summarised in a comprehensive ACM Computing Surveys review [14].

### 2.2. *Input Data (Unstructured Data)*

The data used in this study consists of death and birth event data. A total of 841,126 records were analysed, comprising 449,440 birth records and 391,686 death records. Each record includes a "place of event" field, which can contain location names ranging from the village level to the provincial level, and even specific place names like hospitals.

Given the lack of a standard format for writing the location of an event in this dataset, data cleaning is necessary to reduce data variability. In [15], data cleaning of free-text entries was performed by converting abbreviations, handling letter case, and removing spaces or characters. This approach was also applied to the death and birth vital event data to reduce data variability.

To improve consistency, comprehensive text normalisation was applied using the Pandas and regular expression library in Python. The entire preprocessing workflow was implemented in Python 3.13.5. The cleaning pipeline includes:

- Abbreviation expansion: Using regex patterns, it systematically expanded abbreviations such as 'RS' to 'rumah sakit', 'Kab.' to 'Kabupaten', and 'Kec.' to 'Kecamatan'.
- Standardisation of administrative terms: Custom dictionary was created to strip common administrative descriptors (e.g., 'kota', 'kabupaten', 'kecamatan') using string replacement functions.
- Case and spacing normalisation: All text was converted to lowercase using the str.lower() method, and whitespace was removed using str.replace(' ', '').

This step aligns with best practices in standardising free-form addresses to enable accurate matching, as emphasised in record linkage literature [16]. This was necessary because the data entry for death and birth events did not have clear instructions, leading to variations in how locations were written. Removing spaces, in particular, helped resolve inconsistencies like "Pangkal Pinang" being written with a space in some entries and as "Pangkalpinang" without a space in others. This approach ensured that such variations did not affect the data matching process.

*Master Reference Data (Target Standard)*

The matching process was performed using the master administrative data from MOHA. This dataset was selected because the vital event data also originated from the same ministry, making it more relevant for matching than the statistical working area data from BPS. The master data consists of four administrative levels: province, regency/city, district, and village/urban sub-district.

To facilitate matching, all four levels were combined into a single field. This allowed for a comprehensive matching process where each record was compared against all administrative levels. This approach was necessary because the location information in the death and birth vital event data could be a province, a regency/city, a district, a village, or even a specific location like a hospital name. Similar to the vital event data, the master administrative data was also preprocessed by converting all characters to lowercase and removing all spaces to ensure consistency.

*Multi-Stage Hybrid Standardisation Framework*

This framework operates as a sequential steps, where each stage processes the remaining unstructured entries from the previous stage, progressively reducing ambiguity and increasing standardisation coverage. The entire process was conducted using the Python programming language, with all stages executed within a Jupyter Notebook environment.

*Stage 1: Deterministic Matching*

This initial stage involves direct string comparisons between cleaned CRVS entries and the master reference list. Deterministic linkage provides quick, high-confidence matches but fails when entries contain typos or non-standard formats, a well-recognised limitation in classical linkage workflows [17].

This process was implemented using the Pandas library in Python. Despite its speed and simplicity, this method has limitations. This stage is unable to detect data with typographical errors, preventing them from being matched with the master administrative data. Given that the vital event data for deaths and births is a free-text input, there is a low probability that the entries are free from typos. Additionally, some entries may not be official administrative names but rather local place names, such as hospitals or unofficial neighbourhood names not registered with the MOHA's administrative names.

*Stage 2: Probabilistic Matching*

Remaining unmatched records are paired probabilistically using the RapidFuzz library, specifically leveraging the WRatio function. RapidFuzz was selected due to its superior performance, processing about 40% faster and using memory more efficiently than alternatives such as FuzzyWuzzy, Difflib, Levenshtein, and Jellyfish, especially important for large-scale, multilingual CRVS data [15]. This stage is designed to address typographical errors that occurred during the data entry process. To ensure a high degree of similarity and maintain the assumption of only minor typos, a threshold of 95% was established. This high threshold reduces the risk of false positives by allowing matches only when strings are nearly identical, making it suitable for tasks requiring high precision [18]. This stage employs blocking strategies to reduce the computational burden of pairwise comparisons, aligning with established practices of scalable record linkage [19].

To formalise the scoring mechanism, this study defines a normalised ratio and then applies a composite scoring formula that balances different fuzzy matching strategies.

$$\text{ratio} = \left(1 - \frac{\text{Levenshtein Distance}}{\max(|s1|,\{s2|)}\right) \text{x } 100 \quad (1)$$

$$\text{WRatio}(s1, s2) = \max(\text{ratio}(s1, s2), 0.95 \times \text{token\_sort\_ratio}(s1, s2), 0.9 \times \text{token\_set\_ratio}(s1, s2)) \quad (2)$$

*Stage 3: Geocoding*

For records still unresolved by text-based matching, this study employs geocoding. Geocoding is the process of translating a place name or building into geographic coordinates. The output of this process is a set of coordinates, which are then overlaid onto a map of Indonesia's administrative regions to determine the corresponding regional name (see figure 2).

This study used the Geopy library with data from OpenStreetMap for the geocoding process. The vital event data was fed into the library, which returned a single coordinate point if the location was found in the OpenStreetMap database. If a location was not found, no coordinates were returned.

The resulting coordinates were then overlaid onto an Indonesia GeoJSON map using Python. This method allowed for the retrieval of the administrative region name based on where the point fell within a specific area. The region extracted was the village/urban sub-district level, which facilitates subsequent statistical data creation as it can be easily aggregated to higher administrative levels.

High-volume implementations using hierarchical geocoding frameworks such as TIGER/Line and CASS-style pipelines demonstrate the feasibility of geocoding billions of addresses at scale while maintaining precision [20][21]. Additionally, emerging model-driven frameworks like ELECTRo-map offer an end-to-end probabilistic approach, providing uncertainty-aware geocoding by directly estimating geographic coordinates from text while leveraging contextual cues [22]. Nonetheless, geocoding introduces potential errors when processing ambiguous inputs, such as generic facility names (e.g., "hospital") or entries containing only neighbourhood-level codes like RT/RW, which may result in incorrect spatial assignments [23].

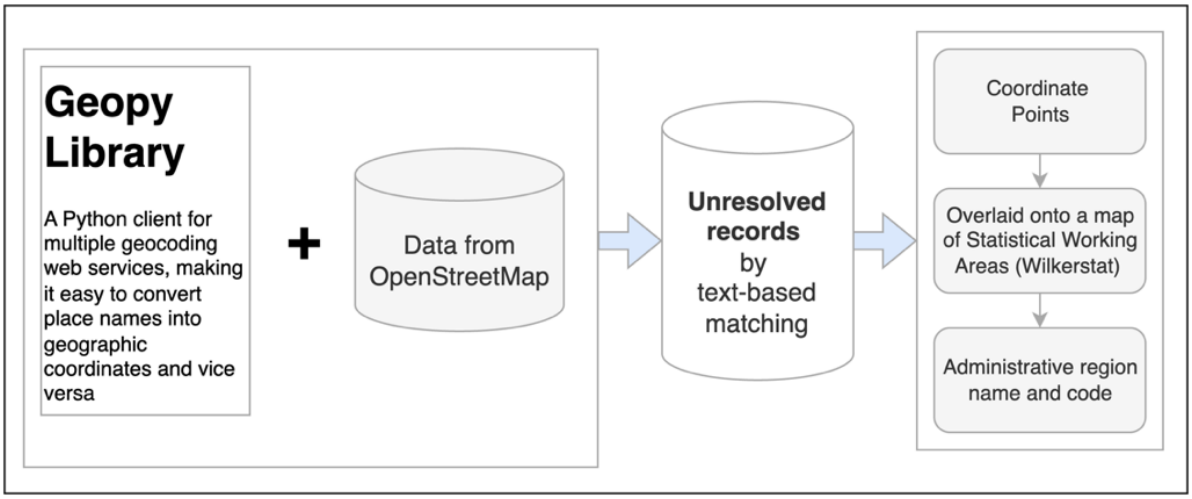


**Figure 58.** Illustration of Geocoding Steps

## 3. Result and Discussion

The multi-stage framework processed 841,126 vital event records through three sequential stages: deterministic matching, probabilistic matching, and geocoding. Each stage was designed to address progressively more complex data quality issues, from exact matches to typographical variations and finally to location ambiguities requiring spatial resolution. The following subsections present the performance of each stage in detail.

Table 1 presents the results of the initial deterministic matching stage, where 29.09% of birth records (130,720 out of 449,440) and 11.83% of death records (46,332 out of 391,686) achieved exact matches with the master administrative data.

**Table 29.** The result of deterministic matching

| Event Type | Match | Not Match | Total | Match Percentage |
|---|---|---|---|---|
| Birth | 130,720 | 318,720 | 449,440 | 29.09% |
| Death | 46,332 | 345,354 | 391,686 | 11.83% |

This stage efficiently solved several common data formatting problems:

- Spacing Issues: Cases like the entry "Banjar Masin" in the death data were successfully matched with "BANJARMASIN" in the master regional data. This highlights the effectiveness of the initial data cleaning process where all spaces and special characters were removed before matching.
- Case Differences: The issue of case sensitivity was resolved by converting all text to lowercase, allowing "Salur Lasengalu" to be matched with "SALUR LASENGALU." This is an important step to ensure consistency and prevent matching failures due to case differences alone.
- Geographical Name Variations: The matching was successful for entries that listed a place name along with its administrative level (e.g., "Kecamatan Ciomas"), which directly matched similar entries in the master data.

Despite its success, the deterministic approach has clear limitations. This stage cannot resolve issues like typographical errors or non-standard naming variations. For example, if "Banjar Masin" was written as "Banjar Masinng," the deterministic method would fail. These failures underscore the necessity of the subsequent probabilistic matching and geocoding stages to handle more complex and varied entries.

The probabilistic matching stage successfully matched an additional 47.89% of the previously unmatched birth records (152,634 out of 318,720) and 22.55% of death records (77,869 out of 345,354) using a fuzzy string similarity threshold of 95% (see table 2). This stage was particularly effective in correcting minor spelling errors, typically involving one or two letters. Despite these corrections, some errors remained due to regional name similarities, underscoring the utility of maintaining a high similarity threshold to minimise false positives.

Some examples of corrected spelling errors at this stage include "TOULIAN OKI," which was successfully matched to the administrative name "TOULIANG OKI," and "PALAWARUKKA," which was matched to "PALLAWA RUKKA." Most of the successful matches were able to correct errors of one or two letters.

However, because some regional names are similar, matching errors still occurred, such as "PAICTAN" being matched to "PAITAN" and "TUGUNIMA" being matched to "TUGU". These cases remain a problem in the probabilistic process, so using a high threshold can minimise these errors.

**Table 30.** The result of probabilistic matching

| Event Type | Match | Not Match | Total | Match Percentage |
|---|---|---|---|---|
| Birth | 152,634 | 166,086 | 318,720 | 47.89% |
| Death | 77,869 | 267,485 | 345,354 | 22.55% |

Cumulatively, the first two matching stages (deterministic and probabilistic) attained a combined match rate of 63.05% for birth records and 31.71% for death records relative to the original dataset as written in Table 32. The remaining unmatched records were processed through geocoding, resulting in successful location assignments for 60.60% of the remaining birth records and 51.85% of the remaining death records (see table 3). Unlike deterministic and probabilistic matching, which rely solely on the input records of births or deaths, geocoding utilises geospatial reference data (in this case, from OpenStreetMap) to pinpoint the exact geographic location of each event. The output of this stage is a pair of latitude and longitude coordinates corresponding to the place of occurrence.

**Table 31.** Geocoding result

| Event Type | Match | Not Match | Total | Match Percentage |
|---|---|---|---|---|
| Birth | 100,651 | 65,435 | 166,086 | 60.60% |
| Death | 138,702 | 128,783 | 267,485 | 51.85% |

The geocoding process successfully mapped several types of ambiguous cases. First, it was able to map building names, such as hospitals, to a coordinate point, which was then assigned to a specific administrative region. Second, it mapped place names that are not officially recorded in the master data. For example, "RS.AR-ROYYAN" could be matched to "KAB. OGAN ILIR" and "Tana Wawo" was matched to "KAB. SIKKA." However, this method also has its drawbacks:

- Vague or Generic Locations: For death/birth event names that only contain an RT or RW number without any other regional information, the matching is likely to be incorrect because there are many similar RT and RW numbers across Indonesia. Similarly, generic information like "RSUD" (General Regional Hospital) cannot be mapped and will result in an incorrect coordinate point.
- Foreign Locations: A location name from another country also cannot be matched in this case, even if the geocoding service provides an accurate coordinate point. This is because the framework is designed to match against Indonesian administrative boundaries.
- Invalid Entries: Lastly, invalid death/birth event names that only contain numbers, dates, or meaningless letters cannot be mapped.

**Table 32.** Cumulative percentage of each stage

| Event Type | Match Percentage (Deterministic) | Match Percentage (Deterministic + Probabilistic) | Match Percentage (Deterministic + Probabilistic + Geocoding) |
|---|---|---|---|
| Birth | 29.09% | 63.05% | 85.44% |
| Death | 11.83% | 31.71% | 67.12% |

## 4. Conclusion

This study proposes a multi-stage hybrid framework that effectively standardises unstructured location data in Civil Registration and Vital Statistics (CRVS) systems by integrating deterministic matching, probabilistic fuzzy matching, geocoding, and spatial proximity analysis. The framework demonstrated significant improvements in match coverage, with a combined matching rate of 85.44% for birth events and 67.12% for death events, as shown in Table 32. These results reflect enhanced accuracy and geographic specificity of vital statistics crucial for public health planning and demographic analyses.

This approach balances computational speed and precision by addressing straightforward cases through rule-based string matching and handling more ambiguous records through similarity scoring and spatial techniques. The layered framework mitigates challenges inherent to free-text entries such as spelling variation, format inconsistency, and ambiguous place names that often plague CRVS datasets, especially in Indonesia's complex archipelagic setting.

While geocoding further resolves ambiguous location assignments, limitations remain, including mapping of generic or incomplete location entries and foreign place names. Future work could investigate integrating transformer-based embedding models for semantic matching, active learning to optimise disambiguation of challenging records, and uncertainty-aware geocoding methodologies to further improve robustness. Such advancements hold promise for producing high-quality, geolocated

vital event data capable of supporting evidence-based policy and enhancing the utility of national CRVS systems.

## References


[1] S. N. L. Mills, "*World - Strengthening CRVS and national ID : January 29, 2016 to October 27, 2017 completion report for the WBG action plan for addressing data gaps in civil registration and vital statistics, 2016-2030 (English)*," World Bank Group, Washington, D.C., 2017.

[2] M. Das, "An overview of vital statistics," *Int. J. Homoeopathic Sci*., vol. 7, no. 2, pp. 519-526, 2023.

[3] V. Velkoff and F. Hollmann, "Uses of Vital Statistics Data," U.S. Census Bureau, Washington, DC, USA, 2008.

[4] Badan Pusat Statistik Indonesia, *Laporan Statistik Hayati Indonesia 2019-2023*, Oct. 2024. [Online].Available:https://www.bps.go.id/id/publication/2024/10/17/f3eaad9790e201d758f8b34c/laporan-statistik-hayati-indonesia-2019-2023.html

[5] S. N. L. Mills, "World - Strengthening CRVS and national ID: Completion report for the WBG action plan for addressing data gaps in civil registration and vital statistics, 2016-2030," World Bank Group, Washington, DC, USA, 2017.

[6] H. Zhang *et al.*, "*FLAP: a framework for linking free-text addresses to the Ordnance Survey Unique Property Reference Number database*," Frontiers in Digital Health, vol. 5, p. 1186208, 2023.

[7] K. Trzos, "Guide to address data standardization – what, how and why?", Algolytics, Jan. 2025. [Online]. Available: https://algolytics.com/guide-to-address-data-standardization-what-how-and-why/

[8] Profisee, "Data Standardization: What It Is and Why It Matters," Aug. 2025. [Online]. Available: https://profisee.com/blog/what-is-data-standardization/

[9] K. Hosseini, F. Nanni, and M. Coll Ardanuy, "*DeezyMatch: A Flexible Deep Learning Approach to Fuzzy String Matching*," in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Online: Association for Computational Linguistics, 2020, pp. 62–69. doi: 10.18653/v1/2020.emnlp-demos.9.

[10] M. Pikies and J. Ali, "*Analysis and safety engineering of fuzzy string matching algorithms*," ISA Transactions, vol. 113, pp. 1–8, Jul. 2021, doi: 10.1016/j.isatra.2020.10.014.

[11] S. Cebeci, M. Özyilmaz, and G. İnce, "*Automatic Standardization System for Free Text Addresses*," in Proceedings of the 27th Signal Processing and Communications Applications Conference (SIU), Sivas, Turkey, Apr. 2019, pp. 1–4, doi: 10.1109/SIU.2019.8806349.

[12] I. Koumarelas, A. Kroschk, C. Mosley, and F. Naumann, "*Experience: Enhancing Address Matching with Geocoding and Similarity Measure Selection*," J. Data Inf. Qual., vol. 10, no. 2, Art. no. 8, pp. 1–16, Jun. 2018, doi: 10.1145/3232852.

[13] D. S., "Mastering text data cleaning in Python: An in-depth guide to preprocessing text data for NLP and machine learning," AIMind, Dec. 3, 2023. [Online]. Available: https://pub.aimind.so/mastering-text-data-cleaning-in-python-an-in-depth-guide-to-preprocessing-text-data-for-nlp-and-1a07539ddb98

[14] X. Hu, Z. Zhou, H. Li, Y. Hu, F. Gu, J. Kersten, H. Fan, and F. Klan, "*Location Reference Recognition from Texts: A Survey and Comparison*," ACM Comput. Surveys, vol. 56, no. 5, pp. 1–37, 2023, doi: 10.1145/3625819.

[15] N. Elmobark, "*A Comparative Analysis of Python Text Matching Libraries: A Multilingual Evaluation of Capabilities, Performance, and Resource Utilization*," International Journal of Environment, Engineering & Education, Vol. 7, No. 1, pp. 48- 60, 2025. https://doi.org/10.55151/ijeedu.v7i1.188.

[16] W. E. Winkler, "*Matching and record linkage*," WIREs Computational Stats, vol. 6, no. 5, pp. 313–325, Sep. 2014, doi: 10.1002/wics.1317.

[17] V. Gupta, M. Gupta, J. Garg and N. Garg, "*Improvement in Semantic Address Matching using Natural Language Processing*," 2021 2nd International Conference for Emerging Technology (INCET), Belagavi, India, 2021, pp. 1-5, doi: 10.1109/INCET51464.2021.9456342.

[18] WinPure, "Common Mistakes In Fuzzy Data Matching," Feb. 11, 2025. [Online]. Available: https://winpure.com/fuzzy-matching-common-mistakes/

[19] C. Nanayakkara, P. Christen, T. Ranbaduge, and E. Garrett, "*Evaluation measure for group-based record linkage*," IJPDS, vol. 4, no. 1, Oct. 2020, doi: 10.23889/ijpds.v4i1.1127.

[20] S. Xu, S. Flexner, and V. Carvalho, "*Geocoding Billions of Addresses: Toward a Spatial Record Linkage System with Big Data*".

[21] P. Christen, A. Willmore, and T. Churches, "*A Probabilistic Geocoding System Utilising a Parcel Based Address File*," in *Data Mining: Theory, Methodology, Techniques, and Applications, Lecture Notes in Computer Science, vol. 3755 LNAI, G. J. Williams and S. J. Simoff, Eds., Berlin, Germany: Springer-Verlag, 2006, pp. 130–145, doi: 10.1007/11677437_11.

[22] B. J. Radford, "*Regressing Location on Text for Probabilistic Geocoding*," in Proc. 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021), Online, Aug. 2021, pp. 53–57, doi: 10.18653/v1/2021.case-1.8.

[23] Z. Yin, C. Zhang, D. W. Goldberg, T. A. Hammond, A. Ma, and X. Li, "*A probabilistic framework for improving reverse geocoding output*," Transactions in GIS, vol. 28, no. 4, pp. 1537–1556, Aug. 2024, doi: 10.1111/tgis.12623.