Fuzzy Name Matching for Untangling Provenance of Colonial Heritage

Sarah Binta Alam Shoilee, Victor de Boer, and Jacco van Ossenbruggen

Vrije Universiteit Amsterdam, the Netherlands
{s.b.a.shoilee, v.de.boer, jacco.van.ossenbruggen}@vu.nl

1 Introduction

Museums grapple with the challenge of managing collections, particularly in determining rightful ownership amidst ethical and legal complexities, especially with limited provenance information. Provenance research, where experts aim to connect objects with their past by examining historical patterns[5], is acknowledged to be time-consuming and expensive. Semantic Web technology and automatic entity linking provides opportunities for scaling up this process, to improve accessibility and provide context[1]. This study, informed by Actor-Network-Theory[3], aims to enhance museum datasets by incorporating collector background information or linking person names across heritage institutions, addressing naming variations, missing attributes, and historical errors. This paper presents a use-case, investigating string matching approaches for entity linking, adapting DeezyMatch[1] for person name matching.

2 Case-Study

In this study, we delve into datasets from two distinct museums. The National Museum of World-culture (NMVW) is an ethnographic museum, traces its historical foundation to colonial times. The **NMVW dataset** (available in Linked Data format), constructed from object meta-data, adheres to the CIDOC-CRM ontology based on Linked Art recommendations[2]. We focus on the "constituents" component, which involves individuals linked to events related to objects, i.e., acquisition or production.

On the other hand, the Museum Bronbeek, originally a veterans' center for Royal Netherlands East Indies Army (KNIL) personnel, provides comprehensive information about military individuals. The **Bronbeek dataset**, with 15,382 constituents, was converted to Linked Data in the context of this research, using the CoW (CSV on the Web) tool[3] The constituent data within the table provides comprehensive information about individuals associated with Bronbeek.

[1] This research is supported by the Pressing Matter project, funded through the NWO National Research Agenda pressingmatter.nl.

[2] https://linked.art/

[3] https://pypi.org/project/cow-csvw/

3 Method

3.1 Initial Exploration

Our investigation into name matching involved 5 string-matching-based approaches before employing more computationally complex methodologies[4]. To explore their efficacy, we use a curated set of 16 correspondences (for 15 different persons) provided to us by museum experts. Table 1 shows the performance of each approach.

Exact String Matching only results in 1 correct match. This low recall can be attributed to character encoding discrepancies, variations in name presentation (including abbreviations and full names within brackets), spelling differences, inclusion of titles in names, and variations in name order.

Initial + Surname Matching approach addresses NMVW's representation convention, utilizing initials followed by full names in brackets and then the surname (e.g., F.W. (Friedrich Wilhelm) Stammeshaus) format. Though recall is improved, issue of character encoding, title variations (e.g., "Baron Haro van Hemert tot Dingshof"), and discrepancies in constructing initials (e.g., "S.J. (Sjoerd) Nauta") persist.

Surname Matching offers flexibility wrt missing name parts and spelling variations, resulting in higher recall, but lower precision Challenges include handling character encoding anomalies and complex surnames(e.g., "Baron Haro van Hemert tot Dingshof", interpreted as "Dingshof").

Fuzzy String Matching is an approximate technique identifying partial matches between text strings, accommodating variations like misspellings, using the "edit distance" metric[4].Despite achieving perfect recall for the 16 identified correspondences, the total retrieved for 16 names is 1883, signaling potential false positives.

Table 1. Performance of rule-based approaches on 15 hand-curated Names (16 known instance correspondence):

	TP	**TP+FP**	**Recall**	**Precision**	**F-score**
Exact Match	1	1	0.0625	1	0.118
Initials Match	10	10	0.625	1	0.769
Surname Match	13	27	0.813	0.482	0.605
Fuzzy Match	16	1883	1	0.009	0.017

[4] Github Repo: `https://github.com/Shoilee/actor_linking`

In summary, the choice of a string matching approach depends on specific objectives and trade-offs between precision and recall. The Initials Match approach demonstrated a good balance, while Surname Match emphasized recall. Fuzzy Match achieved perfect recall but suffered from a high false-positive rate, necessitating precision improvement.

Recognizing the limitations of rule-based approaches, especially in handling evolving datasets and complex naming variations, we additionally investigate a deep learning-based approach.

3.2 DeezyMatch

DeezyMatch [1], an open-source Python library, is designed for advanced string matching and ranking. Initially trained on place names, it uses deep-learning architectural variations with modular hyper-parameter tuning. For person names, we adapted it using the multilingual JRC-Names [2] dataset. The training dataset, "dataset_final_jrc_person.csv," contained positive and negative match data points. Employing the DeezyMatch library, we processed the data, resulting in sets of 3.5 million training examples, 750,000 validation examples, and 750,000 test examples.

Fine-Tuning To refine the model's understanding of NMVW's naming conventions, 6,178 NMVW person instances with Wikidata identifiers[5] were utilized. Conducting federated queries on Wikidata yielded 11,501 alternate name labels. For fine-tuning, positive data points were generated when the correct alternate label exhibited a match ratio above 0.6, while negative samples were created with match ratios below 0.4, aligning with the guidance in [1]. This process resulted in 82,376 data points, enhancing the model through fine-tuning.

4 Evaluation

We evaluate performance on two different alignment tasks using two approaches:

Ground Truth Evaluation We report precision, recall and f-score on the 6,178 instances of the ground truth used for fine-tuning, as presented in the previous section. Results are presented in Table 2. For fine-tuned DeezyMatch , we used 5-fold validation and report rounded results.

Approximate Name Matching To assess person matching performance on our case-study task (NMVW and Bronbeek), 50 correspondences were randomly selected for each strategy's output and manually evaluated on a scale from 0

[5] previously documented in the context of Wikidata:CopyClear initiative: https://www.wikidata.org/wiki/Wikidata:CopyClear/Museum_van_Wereldculturen/Canadian_creators_NMVW_not_in_MNBAQ

(definitely false) to 4 (definitely correct). Precision was calculated, considering any score above 1 as positive, and weighted precision was determined by dividing the sum of scores by the max-total score ($max - score \times \#instance = 200$). The results are presented in Table 3.

Table 2. Recall, Precision and F-score on the Wikidata Ground Truth

	instances	match	correct	Recall	Precision	F-score
Exact Match	6178	3360	3346	0.542	0.995	0.702
Initial + Surname Match	6178	3662	3576	0.579	0.977	0.727
Surname Match	6178	8771	4991	0.808	0.569	0.668
Fuzzy String Match	6178	15028	5515	0.893	0.367	0.520
DeezyMatch	6178	3369	3359	0.544	0.997	0.704
DeezyMatch (after fine-tuning)	1235	673	671	0.544	0.997	0.698

Table 3. Performance evaluation on the NMVW vs Bronbeek by randomly selecting 50 samples

	Total Retrieved	**True Positive from 50 samples**	**Estimated Precision**	**Weighted Precision**	**Estimated Recall**	**Estimated F-score**
Exact Match	351	50	1	0.625	0.030	0.059
Initial + Surname Match	978	46	0.92	0.655	0.08	0.144
Surname Match	51376	9	0.18	0.2	0.80	0.294
Fuzzy String Match	3533880	0	0	0.02	0	0
DeezyMatch	318	49	0.98	0.715	0.03	0.053

5 Discussion and Conclusion

The findings underscore the intricate balance between precision and recall in string matching strategies, urging a meticulous alignment with specific use-case objectives. Controlled strategies may enhance precision but exacerbate recall.

To our surprise, DeezyMatch's performance closely resembles Exact Match, potentially attributed to its strict training criteria, overlooking naming nuances. Training on JRCnames, rich in inter-lingual correspondences but lacking diverse naming conventions, might contribute to this behavior. Random sampling evaluation indicates DeezyMatch's preference for longer strings and less common surnames, reflected in the weighted precision score.

Table 3 reveals challenges even in exact matches, with low weighted precision due to instances having only initials and surnames. This prompts consideration of a human-in-the-loop approach for disambiguation, questioning the suitability of complete automated entity linking in such scenarios.

This study navigates the complexities of person matching in cultural heritage, emphasizing challenges in linking entities solely based on literal names. The data integration approach serves as a foundation for leveraging collector background information in provenance research. The research acknowledges its early stage and focuses on overcoming integration challenges, paving the way for more nuanced exploration in subsequent phases.

References

1. Hosseini, K., Nanni, F., Coll Ardanuy, M.: 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. pp. 62–69. Association for Computational Linguistics (October 2020). https://doi.org/10.18653/v1/2020.emnlp-demos.9, https://aclanthology.org/2020.emnlp-demos.9
2. Jacquet, G., Verile, M.: Jrc-names rdf: Person and organisation spelling variants as found in multilingual news articles. Dataset, European Commission, Joint Research Centre (JRC) (2015), http://data.europa.eu/89h/jrc-emm-jrc-names
3. Latour, B.: Reassembling the social: An introduction to actor-network-theory. Oup Oxford (2007)
4. Navarro, G.: A guided tour to approximate string matching. ACM Comput. Surv. **33**(1), 31–88 (mar 2001). https://doi.org/10.1145/375360.375365, https://doi.org/10.1145/375360.375365
5. Tompkins, A.: Provenance Research Today. Lund Humphries (2021)