A TWO-TIRE APPROACH FOR ORGANIZATION NAME ENTITY RESOLUTION

Almuth Müller[1] [a], Achim Kuwertz[1]

[1]*Fraunhofer IOSB, Fraunhoferstraße 1, 76131 Karlsruhe, Germany*
almuth.mueller@iosb.fraunhofer.de, achim.kuwertz@iosb.fraunhofer.de

Keywords: Entity Resolution, Record Linkage, Deduplication, Natural Language Processing, Fuzzy Matching

Abstract: This paper presents a concept for a two-tire semi-automated approach for business data entity resolution. Resolving entity names is generally relevant e.g. in business intelligence. When applied, several difficulties have to be considered, such as name deviations for an organization. Here, two types of deviations can be distinguished. First, names can differ due to typos, native special characters or transformation errors. Second, an organization name can change due to outdated designations or being given in another language. A further aspect is data sovereignty. Analyzed data sources can be under direct control, e.g. in own data storage systems, and thus be kept clean. Yet, other sources of relevant data may only be publicly available. It is in general not recommended to copy such data, due to e.g. its amount and data duplication issues. The proposed two-tire approach for entity resolution thus not only considers different kinds of name derivations, but also data sovereignty issues. Being still work in progress, it yet has the potential to reduce the effort required when compared to manual approaches and can possibly be applied in different areas where there is a significant need for harmonized data and externally curated systems are not feasible.

1 INTRODUCTION

Data quality management is a vital aspect of the data management process. It is a very broad field by itself, with data harmonization playing a significant role. Data harmonization refers to all efforts combining information from diverse sources and providing analysts with a comparable representation. This is becoming increasingly important in today's data-driven world, where data is frequently distributed across multiple data sources.

There are many different fields of application, like data warehouses with extract-transform-load (ETL) processes, Internet-of-Things (IoT) applications where data is used for machine learning or the healthcare domain where patient data is distributed across different practitioners.

In most cases, non-harmonised data entries result in an inaccurate analysis. This inevitably has a negative impact on the development of models, forecasts, or simulations, leading to a reduction in the analysis procedure's reliability, user acceptance, and user sat-

isfaction. Therefore, data harmonization is of interest to both theoreticians and practitioners.

To harmonize data from different sources and derive meaningful insights, several activities must be considered. This study focuses on entity recognition (ER), deduplication, and record linkage (RL).

Record linkage (RL) describes the task of cleaning and joining different representations of the same real-world entity across different datasets (Fellegi and Sunter, 1969, Winkler, 1990, Mirylenka et al., 2019). Entity resolution (ER) and deduplication ensure that a real-world object is represented by just one single record (Elmagarmid et al., 2007, Binette and Steorts, 2022). When record linking, entity resolution, and deduplication are applied, a single representation of an entity can be derived that is enriched with information originally spread across different datasets.

In recent years, significant progress in data harmonization has been accomplished, primarily through data mining and machine learning (Gottapu et al., 2016, Mudgal et al., 2018, Li et al., 2021). Despite the fact that there are numerous commercial systems available, most of them are "black box" systems from the user's perspective. Often, very superficial information is available about

---

the methods used. As a result, such systems do not allow direct insight into the quality of their results, which would be particularly important in science and healthcare. Besides their high price, this is another reason why such systems are not feasible for some areas of application. Furthermore, many of the systems are focused on specific domains with unique data requirements. Due to the wide range of applications, none of the existing systems can be used universally or without adaptation to a new application domain (Köpcke et al., 2010).

This paper is based on a business intelligence use case for research institutes. As a result, important data sources include project partners, patents, and paper publications. Considering the aforementioned issues, existing systems could not be adequately adapted.

Instead, a new two-tiered approach is proposed, based on the latest research on the relevant tasks (ER, RL, etc.) and tailored to the circumstances and requirements of the use case under consideration.

The paper presents a conceptual architecture with recommendations of available or experimental methods for each step of the process. The implementation of those methods is currently ongoing and will be presented as future work. The considered use cases includes datasets with different data schemes and data sovereignty.

Chapter 2 describes the use case on which the paper is based and the properties of the data sources considered. In chapter 3, the developed concept for the recognition of organizations across datasets is presented. The concept addresses the aspects of data sovereignty by introducing a catalog of name variations, the Corporation Catalog, as a central part of the concept. Chapter 4 then describes the construction of the Corporation Catalog, along with methods for resolving the different name variations by which organizations may be represented. To recognize records that belong to the same organization Natural Language Processing (NLP) and fuzzy logic-based techniques are investigated. Work in progress is discussed in chapter 5. The paper concludes with a summary in chapter 6.

The principles mentioned in this paper can be transferred to other areas that face similar conditions.

2 USE CASE

The work presented in this paper is based on a need to gain insights into an organization and its research partnerships in order to conduct targeted research. The relevant data about project participations, patents and publication is spread among different datasets, both locally and externally stored.

The individual entries do not necessarily have to be duplicates, in the sense of double, identical entries. More often an entity can have different occurrences across multiple data sources, e.g. listing several patents of an institute in one dataset and the project participation of the same institute in another dataset.

A concept for harmonizing this data must therefore be able to deal with three main challenges:

- own and third-party data sovereignty,
- various data schemes with few common features, and
- spelling mistakes or variations of organization designations

The challenges are discussed in more detail below.

2.1 Data Sovereignty

Data sovereignty is an important but often overlooked factor. Some of the relevant data sources for analysis are stored locally in the company's data management system. Other relevant data sources can be obtained from externally accessible databases. Diverse sources of data, including data provided outside the organization, can lead to significant insights into new lines of research (Boscoe et al., 2011).

While locally stored records can be edited directly and kept clean, this is not feasible with external ones. Due to the volume of such data sets, it is not recommended to copy external data sources into the organization's data storage system. Especially since this would cause additional problems due to double data storage.

This shows that a concept for data harmonization must follow different approaches depending on the underlying data sovereignty.

2.2 Data Schemes

Another challenge to the harmonization process arises from the very different data schemes of the various data sources. The various data schemes are due to different natures of the data sets, such as patent databases and project or publication databases.

The only common field among the data schemes under consideration is usually only the name feature, which contains the designation of an organization. It therefore acts as a "primary key" across multiple records. Therefore, especially when investigating suitable methods for deduplication and record linkage, the focus for recognizing an organization in a dataset is based on the organization designation.

Without the benefit of multiple feature columns, less information is available and fewer approaches are feasible (Kaufman and Klevs, 2021).

2.3 Spelling Mistakes and Deviations

An organization's designation is prone to misspellings and other types of deviations. A conceptual distinction can be made between two types of such name deviations.

Spelling mistakes can occur due to e.g. typos, native special characters or transformation errors (Hernández and Stolfo, 1998). Spelling mistakes lead to discrepancies in the designation at the letter level.

Besides that, designations of an organization can show major deviations. Organizations, for example, frequently have an official English name in addition to their domestic one. Furthermore, organizations' designations may have changed throughout time e.g. Daimler Benz, Daimler Chrysler, and Mercedes. In addition, organizations can have subsidiaries that are to be added to the parent company for the analysis. Those deviations can lead to completely different designations.

Both groups differ significantly in the methods that can be used to identify them. It is not possible to handle this task with a single tool, but a more comprehensive approach to deal with such different representations of organizations is required.

3 CONCEPTUAL TWO-TIRE APPROACH

The conceptual architecture for semi-automated data curation approach for business data derives from two main challenges: data sovereignty and the two significant groups of name deviations.

Externally managed data cannot be changed directly and stored in a harmonized manner. In general, this can be circumvented by maintaining local lookup tables referencing IDs in these external records and flagging duplicates or records from the same organization.

This paper presents a different approach that offers the benefit of a feature store in addition to the same ease of use as linking to external record IDs. This approach enables a more efficient harmonization process in the long term. Instead of cleaning up errors in organization name features, the presented concept aims to collect the name deviations of organizations and store them grouped by organization, as shown in figure 1.
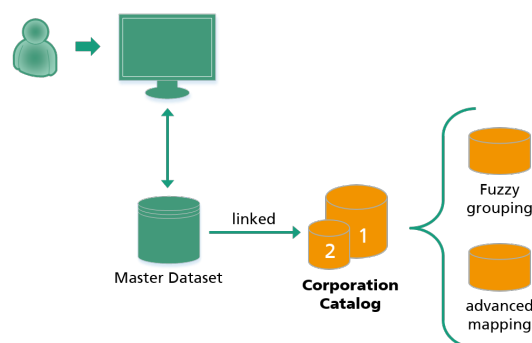

Figure 1: The Corporation Catalog itself consists of two components. The first component groups organization names with fuzzy deviations, e.g. due to typos or phonetic shifts. The second component identifies more complex name variations like different languages or historic organization names

The collection of such groups of name deviations is referred to as the Corporation Catalog in this paper. The Corporation Catalog collects, groups, and manages the name deviations according to the underlying organization.

The Corporation Catalog contains two components that are build in two consecutive steps, as shown in figure 1. Those components correlate with the two groups of name deviations. The first component deals with the spelling mistakes at the letter level. This component is therefore referred to as fuzzy grouping. The second component deals with the more complex name deviations, like names in different languages and is therefore referred to as advanced mapping.

It can be beneficial to keep the result of the second step separate from the results of the first, especially when it comes to parent and subsidiary organizations. Users may want to consider them independently for some analysis. The concept presented, therefore, suggests storing the result of the second step as an additional component in the Corporation Catalog. In chapter 4, the construction of the Corporation Catalog is explained in greater detail.

Technically, the Corporation Catalog represents another dataset in the organizations data storage system. Each group in the Corporation Catalog can be accessed using a unique identifier. The concept works best if the identifier can be linked to a master data record in which the official names of the organizations are stored.

This master data record then represents an interface. The Corporation Catalog can be used through this interface to harmonize data. There are two types of use, depending on the data sovereignty of the target database.
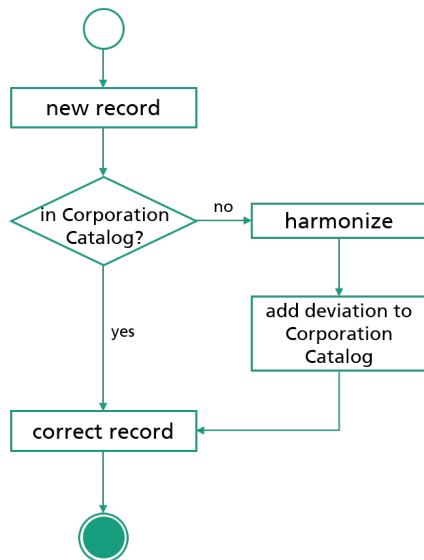
Figure 2: This flowchart shows how the Corporation Catalog helps harmonize records stored locally in an organization's data storage system. New data records with common deviations may already be known to the Corporation Catalog and can therefore be corrected directly. The harmonization pipeline processes unknown deviations only. The Corporation Catalog grows continuously as new deviations are fed back into it.

3.1 Locally Stored Data

Assuming a Corporation Catalog with multiple spelling errors and name variations for distinct organizations is present. The Corporation Catalog supports and optimizes the harmonization of datasets stored locally in the organization's data storage system in this scenario.

Some organizations' designation show typical name deviations that appear in many data sources. With the classic procedure of cleaning each dataset separately, such typical deviations would have to be repeatedly identified and processed. In the case of the Corporation Catalog, this repetitive work can be reduced. The flowchart in figure 2 illustrates the process. By comparing the characteristics in the dataset with those in the Corporation Catalog, several entries for one organization can easily be found and harmonized. The remaining entries can then be further processed using harmonization methods, described in chapter 4.

The Corporation Catalog's database is always increasing since newly discovered name deviations are submitted back into it. This increases the potential of the Corporation Catalog. The Corporation Catalog collects a growing number of name deviations throughout time. This postpones the need to apply the harmonization pipeline to clean up new items in a
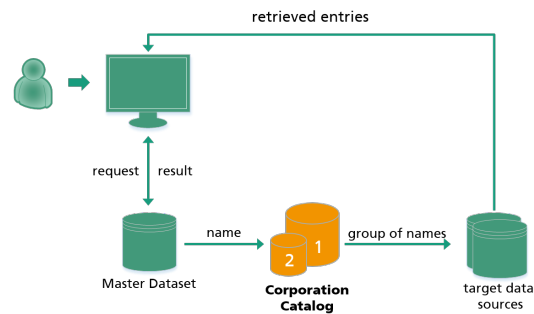

Figure 3: The Corporation Catalog is used to perform an on-the-fly data cleaning and linking on externally stored data sources. Via a master dataset a user starts a query based on the organizations gold standard name. The Corporation Catalog supplements this name with the available deviations and forwards an advanced search to the target data sources. The retrieved entries are displayed to the user.

growing database. There is a certain probability that the newly added entries are already known to the Corporation Catalog.

3.2 External Data

In addition to the resource-saving cleaning of local datasets, the Corporation Catalog enables the possibility of viewing external data sources in a similarly cleaned manner. External datasets do not allow simple, persistent changes to the data. The Corporation Catalog can be interposed to enable appropriately cleaned queries on these data sources as well.

Such a query using the Corporation Catalog is shown schematically in figure 3. The interface between the user and the target data source is the master dataset. The master dataset contains the gold standard names of the organization. A user can use this master dataset as an interface to send queries to other data sources.

The query is forwarded to the Corporation Catalog and the corresponding group in the Corporation Catalog is selected. The user's query is modified to include the group's name deviations, e.g. query chain with logical OR. After that, the modified query is sent to the target database. The target database's entries are collected and delivered in response to the user's request. This method employs on-the-fly processing. The user receives the result directly from the target database.

Since this procedure might be associated with increased computing power, it should only be used where other options are impossible.

4 CONSTRUCTION OF THE CORPORATION CATALOG

As described in chapter 2.3, name deviations of organizations in datasets are mainly due to spelling errors, outdated designations or the use of different languages.

A two-tier approach to constructing the Corporation Catalog in two steps is proposed to handle those name deviations. Figure 1 depicts the two components that will be discussed here.

4.1 Component One: Fuzzy Groups

The first component of the Corporation Catalog is dealing with an organization's "fuzzy records." Because the names deviate at the letter level due to typos or phonetic shifts, these records cannot be directly linked to a real-world organization.

A group of algorithms that specializes in matching those deviations is called fuzzy string-matching algorithms (Filipov and Varbanov, 2020). These algorithms calculate the similarity of expressions and provide a probability that two expressions are identical, minus the errors. Some of the common fuzzy matching algorithms are e.g. Levenshtein Distance (or Edit Distance), Damerau-Levenshtein Distance, Jaro-Winkler Distance and Jaccard Index (Jaccard, 1912, Damerau, 1964, Levenshtein, 1965, Winkler, 1990, Bard, 2007).

Usually, a fuzzy matching algorithm alone does not provide the best match. As a result, it is common to use multiple algorithms. The outputs of the individual algorithms are added together to obtain a final result. The individual results can be provided with weights that reflect the reliability of one algorithm for the specific application (Wang et al., 2019, Gregg and Eder, 2022).

As a result, these methods can be used in machine learning applications. For example, a system can be trained to find the best weights for the individual fuzzy methods. An implementation for this component of the Corporation Catalog is currently being evaluated. The Python library Dedupe (Gregg and Eder, 2022) is analyzed for its usability for the discussed use case. First results of the evaluation are discussed in chapter 5

The result of the first component, processing only the fuzzy name deviations, will lead to several groups for one organization. Those groups could be one group per language of the organizations name, for example. This is why a second component is relevant. The advanced mapping collects these individual groups and combines them into a more unified group.

This new group then can be linked to the corresponding entry from the master database.

4.2 Component Two: Human Machine Interaction for Advanced Mapping

Multiple records for one organization can result from foreign-language names, outdated names or hierarchical company dependencies (see chapter 2.3). Such entries cannot be grouped using fuzzy matching, as they often differ completely from one another. Therefore, for some organization, the fuzzy grouping of the Corporation Catalog will contain multiple groups, e.g. a group with the German name deviations and another group with the English name deviations. The task of the second component is to identify and merge these fuzzy-groups.

It is currently not reasonable to assume that these fuzzy groups can be fully automatically harmonized. It is very likely that a combination of human input and machine assistance will be required.

For the extended mapping to provide satisfactory results under this premise, the performance of the fuzzy component must meet certain requirements.

It is generally much easier for humans to handle elements that are correctly grouped. Searching within one group for a small amount of wrongly added elements is a very tedious task to be done manually. Especially since the incorrect elements do show a certain similarity to the correct elements, otherwise they would not have been grouped by the fuzzy matching algorithms. It is also quite easy for humans to identify connections between groups that actually belong together.

Therefore, it is feasible for the first component to output several fuzzy-groups for one organization. The fuzzy-groups do not have to be complete and contain all entries of one organization. On the other hand, a fuzzy-group must contain predominantly correctly grouped entries. As a result, each group does not have to be checked for incorrect entries in the second component. It is then sufficient for the advanced mapping to look at the groups superficially and to further unite them accordingly.

At the current time, no method was identified that could reliably recognize all the very different name deviations. There are several interesting methods to support a human-in-the-loop approach for recognizing and combining the different groups. Methods for machine translation appear to be a promising approach to identifying groups with foreign languages (Xu et al., 2021). However, it still needs to be specifically investigated how such a procedure can be integrated and how much training effort is required. Espe-

cially as organization names might not be direct translations. Embeddings and word vector similarity metrics could be used to identify historical designations or organization hierarchies (Mohammadkhani, 2020, Chen et al., 2019, Obraczka et al., 2021).

Knowledge graphs are another option. In a graph-like structure, knowledge graphs integrate entities with properties and relationships, as well as accompanying metadata about entity and connection types. DBpedia (Auer et al., 2007) is an example of a general knowledge database. There are also some knowledge databases, especially for organizations, e.g. Virtual International Authority File (VIAF). Alternative organization names or dependencies for parent and subsidiary organizations can be stored in knowledge databases. Groups belonging to the same organization can be recognized in the Corporation Catalog by extracting this information. Most publicly available knowledge databases include interfaces for retrieving data in a targeted manner.

5 DISCUSSION AND FUTURE WORK

Data harmonization to ensure data quality is an important part of an overall data management strategy. Various actions are required, including initial one-time actions and ongoing activities to maintain data quality in a growing database. The presented conceptual architecture addresses both actions. The Corporation Catalog supports initial cleaning of new datasets as well as a repetitive cleaning of new data records. The feature-memory effect of the Corporation Catalog also decreases the workload of the ongoing data cleaning activities.

An implementation for the first component of the Corporation Catalog is currently in progress. The use of the Python library Dedupe (Gregg and Eder, 2022) is analyzed for its suitability for the discussed use case. Access to training data poses a challenge here, which is why the use of synthetically generated datasets is being investigated.

The Corporation Catalog must meet the two requirements of homogeneity and completeness regarding the fuzzy groups, as mentioned in chapter 4.2. With appropriate machine support, the user can further combine the groupings in the second stage. The degree of homogeneity and completeness of the fuzzy groups are used to assess if these requirements have been met.

The values of both dimensions are represented as bar charts in figure 4. Dedupe creates groups in which 85 percent of the entries are correct. As a result, the
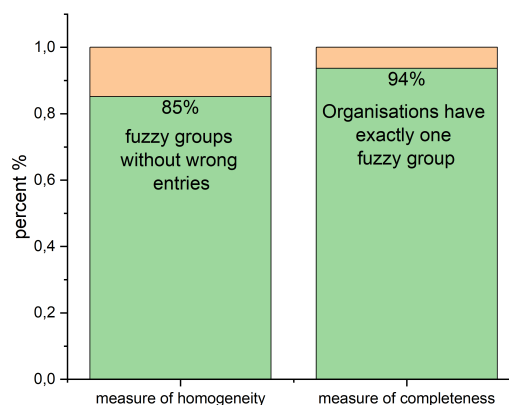

Figure 4: The degree of homogeneity and completeness for the fuzzy groups is shown. The degree of homogeneity indicates what percentage of the groups formed are free of incorrect entries. The degree of completeness indicates what percentage of the corporations are assigned to just one group.

homogeneity of the established groups can be viewed as good. Dedupe is meeting the use cases criteria for homogeneity.

Completeness has an even higher value; 94 percent of the corporations are represented by just one group. It should be noted that the synthetic datasets only reflect differences in designations caused by spelling errors. The high value of completeness means that Dedupe can assign the names resulting from spelling errors to the same group. This is important because, in the second stage, the user can then focus on groups of alternative designations for a corporation.

While it is common for deduplication and record linking methods to require some level of data sanitation as preprocessing, this may not have the desired impact on the record linking result (Randall et al., 2013). Therefore, the approach in this paper wants to keep preprocessing to a minimum. Typical deviations, such as umlauts and upper and lower case letters are automatically corrected in advance. The rest should be done in the course of the harmonization process, since the different datasets generally have large differences in possible cleanup steps. The feasibility of this approach still needs to be checked in the future. Instead of the currently used synthetic datasets, real annotated datasets are best suited for this.

For the second component, current research approaches were presented in chapter 4.2, which the authors of this paper consider promising. A deeper evaluation of the approaches is necessary. However, the prerequisite for an evaluation regarding the use in the application presented here is the performance result of the first component.

In addition to the name feature, the Corporation Catalog could also store other feature values, which help to identify organizations in datasets.

Although the paper focused on business data, the problem of duplicate or unrecognized entries for the same entities is evident in all types of stored data. The application of the presented concept should thus be checked for other areas of application and investigated accordingly.

6 CONCLUSION

A semi-automatic two-tiered approach for record linkage and entity resolution of business data was presented in this paper. The approach considers the data sovereignty of different data sets as well as different reasons for organization name variations. The topic of data sovereignty is often neglected in current research, although it is becoming increasingly important in practice.

The paper presents a conceptual architecture. The chapters 4.1 and 4.2 discuss recommendations of available or experimental methods for each step of the conceptual process. The implementation of these methods is currently in progress and as such has been discussed in chapter 5.

This approach is specially designed for the case where only the company name is available as a data set spanning feature for deduplication and entity resolution of organization data. Therefore, a two-tire approach was considered using fuzzy logic and NLP-based deep learning techniques.

The first component of the approach is designed to handle character-based name variations and thus fuzzy logic-based techniques can be used. First results show that homogeneous and complete groups, regarding name deviations at the letter level, can be achieved.

The second component of the approach deals with more complex deviations that have few similarities. While the first component can be fully automated, the second requires human-machine interaction.

Overall this approach has the potential to reduce the effort required compared to a mostly entirely manual data curation. In addition, the use of computationally expensive record linkage and entity resolution methods can be minimized by using the Corporation Catalog. The potential of such an approach could be realized in different areas where there is a significant need for harmonized data and externally curated systems are not feasible.

REFERENCES

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., and Ives, Z. (2007). DBpedia: A Nucleus for a Web of Open Data. In Aberer, K., Choi, K.-S., Noy, N., Allemang, D., Lee, K.-I., Nixon, L., Golbeck, J., Mika, P., Maynard, D., Mizoguchi, R., Schreiber, G., and Cudré-Mauroux, P., editors, *The Semantic Web*, Lecture Notes in Computer Science, pages 722–735, Berlin, Heidelberg. Springer.

Bard, G. V. (2007). Spelling-error tolerant, order-independent pass-phrases via the damerau-levenshtein string-edit distance metric. In *Proceedings of the Fifth Australasian Symposium on ACSW Frontiers - Volume 68*, ACSW '07, pages 117–124, AUS. Australian Computer Society, Inc.

Binette, O. and Steorts, R. C. (2022). (Almost) All of Entity Resolution. *arXiv*.

Boscoe, F. P., Schrag, D., Chen, K., Roohan, P. J., and Schymura, M. J. (2011). Building Capacity to Assess Cancer Care in the Medicaid Population in New York State. *Health Services Research*, 46(3):805–820.

Chen, X., Campero Durand, G., Zoun, R., Broneske, D., Li, Y., and Saake, G. (2019). The Best of Both Worlds: Combining Hand-Tuned and Word-Embedding-Based Similarity Measures for Entity Resolution. In Grust, T., Naumann, F., Böhm, A., Lehner, W., Härder, T., Rahm, E., Heuer, A., Klettke, M., and Meyer, H., editors, *BTW 2019*, pages 215–224. Gesellschaft für Informatik, Bonn.

Damerau, F. J. (1964). A technique for computer detection and correction of spelling errors. *Communications of the ACM*, 7(3):171–176.

Elmagarmid, A. K., Ipeirotis, P. G., and Verykios, V. S. (2007). Duplicate Record Detection: A Survey. *IEEE Transactions on Knowledge and Data Engineering*, 19(1):1–16.

Fellegi, I. P. and Sunter, A. B. (1969). A Theory for Record Linkage. *Journal of the American Statistical Association*, 64(328):1183–1210.

Filipov, L. and Varbanov, Z. (2020). On Fuzzy Matching of Strings. *Serdica Journal of Computing*, 13(1-2):71–80.

Gottapu, R. D., Dagli, C., and Ali, B. (2016). Entity Resolution Using Convolutional Neural Network. *Procedia Computer Science*, 95:153–158.

Gregg, F. and Eder, D. (2022). Dedupe.

Hernández, M. A. and Stolfo, S. J. (1998). Real-world Data is Dirty: Data Cleansing and The Merge/Purge Problem. *Data Mining and Knowledge Discovery*, 2(1):9–37.

Jaccard, P. (1912). The Distribution of the Flora in the Alpine Zone.1. *New Phytologist*, 11(2):37–50.

Kaufman, A. R. and Klevs, A. (2021). Adaptive Fuzzy String Matching: How to Merge Datasets with Only One (Messy) Identifying Field. *Political Analysis*, pages 1–7.

Köpcke, H., Thor, A., and Rahm, E. (2010). Evaluation of entity resolution approaches on real-world match

problems. *Proceedings of the VLDB Endowment*, 3(1-2):484–493.

Levenshtein, V. I. (1965). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics. Doklady*, 10:707–710.

Li, B., Miao, Y., Wang, Y., Sun, Y., and Wang, W. (2021). Improving the Efficiency and Effectiveness for BERT-based Entity Resolution. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13226–13233.

Mirylenka, K., Scotton, P., Miksovic, C., and Alaoui, S.-E. B. (2019). Linking IT product records. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 101–111. Springer.

Mohammadkhani, M. (2020). A Comparative Evaluation of Deep Learning based Transformers for Entity Resolution. Master's thesis, Otto-von-Guericke-University, Magdeburg.

Mudgal, S., Li, H., Rekatsinas, T., Doan, A., Park, Y., Krishnan, G., Deep, R., Arcaute, E., and Raghavendra, V. (2018). Deep Learning for Entity Matching: A Design Space Exploration. In *Proceedings of the 2018 International Conference on Management of Data*, SIGMOD '18, pages 19–34, New York, NY, USA. Association for Computing Machinery.

Obraczka, D., Schuchart, J., and Rahm, E. (2021). EAGER: Embedding-Assisted Entity Resolution for Knowledge Graphs. *arXiv*.

Randall, S. M., Ferrante, A. M., Boyd, J. H., and Semmens, J. B. (2013). The effect of data cleaning on record linkage quality. *BMC medical informatics and decision making*, 13:64.

Wang, J., Lin, C., and Zaniolo, C. (2019). MF-Join: Efficient Fuzzy String Similarity Join with Multi-level Filtering. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, pages 386–397.

Winkler, W. E. (1990). String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. *Proceedings of the Section on Survey Research Methods*, page 9.

Xu, H., Van Durme, B., and Murray, K. (2021). BERT, mBERT, or BiBERT? A Study on Contextualized Embeddings for Neural Machine Translation. *arXiv*.