

Record Linkage for Historical German VET Data: Towards Linked Labor Market Data

Thomas Reiser ¹, Jens Dörpinghaus ^{1,2,3}, and Petra Steiner²

Abstract: Digital archives often describe their catalogs with digital objects where each record represents a physical document that can be found in the archive. While this approach is most effective to describe data that is present in the archive, there are some archives where the absence of missing documents can impact its quality. To approach methods that can be used to assess the completeness of an archive, we use record linkage methods on the example of the occupations archive at the Federal Institute for Vocational Education and Training (BIBB), combined with the genealogy of vocational education. Our idea is that every record from the genealogy – which describes the history of vocational training in Germany – should have a matching record in the occupations archive – which contains the legal regulations that build the foundation for the genealogy – which implies a mapping between the two data sets. To this end, we create pairs of potential matches between records of both data sets and separate them in two classes: matches and non-matches. These pairs are then used to train different classifiers that should be able to categorize unseen record pairs. The results show that the selected classifiers give reasonable results but need some more improvements for better reliability.

Keywords: record linkage, labor market data, classification

1 Introduction

Digital archives provide a great opportunity when it comes to the automated processing of archival records. Structured standards like METS/MODS, ISAD(G), and DCMI provide great interoperability between records from different sources. However, human errors like typos or misrecognized information from the physical documents can still make it into data sets. This makes the task of mapping archival records from different sources even more challenging.

The occupations archive at the Federal Institute for Vocational Education and Training (BIBB) in Germany contains more than 7,000 historical training regulations from 100 years of German history. However, many documents in the archive had to be disposed due to damages that were too severe for any restoration and some documents might have never made it into the archive stock in the first place, but could be found in other archives

¹ University of Koblenz, Germany, treiser@uni-koblenz.de,  <https://orcid.org/0009-0007-5452-7800>;
doerpinghaus@uni-koblenz.de,  <https://orcid.org/0000-0003-0245-7752>

² Federal Institute for Vocational Education and Training (BIBB), Bonn, Germany,
doerpinghaus@uni-koblenz.de,  <https://orcid.org/0000-0003-0245-7752>; steiner@bibb.de

³ Linnaeus University, Department of Computer Science and Media Technology, Växjö, Sweden,
doerpinghaus@uni-koblenz.de,  <https://orcid.org/0000-0003-0245-7752>

that provide digital catalogs. Although completeness of an archive is nearly impossible to guarantee, the quality of the archive might be improved by identifying gaps in it.

Historical data is not necessarily given in form of archives, but also other structured formats like tables or linked data. The genealogy of vocational education is a tabular data set containing a list of historically state-recognized occupations in Germany. Although it has not many exactly matching fields to archival data, in the case of the occupations archive, there are some commonalities, like the occupation name that is also included in the titles, which could be sufficient for a successful matching of record pairs.

A successful mapping of these two data sets, but also a matching to records of other archives, should help identifying gaps in the occupations archive in future research. The research questions guiding this research are:

1. How can record linkage methods be applied on data sets of historical labor market data?
2. Which are the most efficient classifiers for the linkage of historical labor market data?

To answer these questions, we apply a small selection of record linkage methods on a labeled subset of pairs from records of the occupations archive and the genealogy of vocational education, containing 310 matches and 164 non-matches. We aim to provide a proof of concept that shows the potential of record linkage methods in labor market research to pave the way for future research that might include data sets that were collected from open available digital archives that could contain records that are relevant for the occupations archive as well.

2 Related Work

Data matching, also referred to as record linkage, is a topic of significant interest within many fields, e.g., medical research, and other areas that contain personal data where entities from different sources might refer to the same person. However, the number of publications directly addressing this subject is limited. For further information, readers are directed the works of Christen [Ch12] or [St05]. It bears a close relation to a multitude of other problems, including the entity resolution (ER) problem, which is defined as the process of determining if multiple data records refer to the same entity in the real world [OI25], as well as product matching [Al24] and the privacy and data protection of data [Sc07]. This issue can be regarded as a subproblem of data fusion in general, see [LKC06]. Given the generally heterogeneous nature of data, there is a plethora of examples in the extant literature. These include spatial data (see, e.g., [OM08; WF99]), historical data, see [JSS02], healthcare data (see, e.g., [Co12]), and life sciences data, see [Gu22; Pa07].

In light of the paramount importance of linking and matching data, it is imperative to delve into the annals of data processing. This concept has been delineated since as early as 1974,

as evidenced by the seminal works of Okner (1974, [Ok74]) and [Ca77; Ka75]. The process of data matching and linking is typically executed through the utilization of similarity functions. Subsequently, either artificial intelligence (AI) or distance methods are employed to analyze and establish relationships between the data points [Da07].

3 Data Sets

The two data sets used for our research are the archival records from the occupations archive and the genealogy of vocational education. They refer to two different types of entities, but are still related: While the archive data describe documents in the occupations archive that are training regulations, the genealogy only focuses on which occupation was recognized by the German state, including the time period where this was the case. However, as the training regulations are the legal foundation for when a training had state recognition, there are connections between the two data sets.

As both data sets were created by humans, there might be smaller inconsistencies between them, e.g., through typos, missing data, and spelling inconsistencies. These are the challenges that are commonly addressed by record linkage approaches.

Nevertheless, there are some additional challenges specific to data sets like ours. Some occupations for example are only given in the form of one gender, while others are at times given with both, which makes a difference in German language. For example, one record could be *Kaufmann im Einzelhandel* / *Kauffrau im Einzelhandel* (retail salesman / retail saleswoman) in the genealogy, while a matching record in the archive could refer to *Einzelhandelskaufmann* (retail salesman). Not only does the first version give the occupation title with two genders, but the separate words are also combined into a compound word in the archive version.

4 Method

Our method is based on the *Data Matching* book by Peter Christen [Ch12] and utilizes the *RecordLinkage* Python library [De19]. It performs the five steps that are found in both of these sources: preprocessing, indexing, comparing, classification and evaluation. The key idea of our method is the creation of a set of potential matches and classifying them into two clusters: matches and non-matches.

4.1 Preprocessing

Although data matching techniques are able to handle inconsistencies between records to some extent, preprocessing of the data is crucial. For our data sets, this includes cleaning

string data, in particular titles from the records of the occupations archive, but also the occupation names that should be found in both the genealogy of vocational education and the occupations archive. Because the titles are considered to contain the occupation's name that is regulated in the underlying document, both the title and occupation name are key columns in the archive data set. This string data is cleaned by: 1. converting all characters to lowercase, 2. removing special characters and punctuation, 3. removing multiple space characters in a row which can exist due to human errors when creating the records, and 4. removing umlaut characters. Doing so helps removing discrepancies between records that refer to the same entity.

Additionally, several dates can be found in the records of the occupations archive. These dates are: date of issue, entry into force, and year of publication. The genealogy of vocational education on the other hand contains one date column, when the record's occupation was created or revised. In its current status, the date column of the occupations archive's data set are not fully harmonized and also contain string information in some records. For this, the date columns of the archive data set are considered as string and a method to extract dates from string is applied to obtain data that is formatted as date.

On the other hand, the date for some of the documents is not known due to incompleteness or damages on some of the physical documents. As the date is a key feature to match the records as the same occupation can exist in multiple versions in the genealogy, techniques to fill the date gaps in the archive data set like using the mode date are not appropriate. Therefore, records that contain no information on the date in any of the three aforementioned date columns are dropped for the sake of this research.

A small excerpt of the preprocessed archive data set is shown in Table 1.

Titel	Titelzusatz	Beruf	Inkraft-treten	Erscheinungs-jahr	Erlassdatum
Bekannt-machung der Fortbil-dungspro-file zu den ...		Abwasser-meister, Fachkauf-mann für Außen-wirtschaft ...	NaT	2006-05-21	2006-02-14
Prüfungsan-forderungen für Achat-bohrer	Industrie-Facharbeiter-prüfungen	Achatbohrer	NaT	NaT	1937-01-04

Tab. 1: Excerpt of the preprocessed table of the archive data set. NaT means that no date was found in a column for the particular record. *Titel* is title, *Titelzusatz* is title suffix, *Beruf* is occupation, *Inkrafttreten* is entry into force, *Erscheinungsjahr* is release year, *Erlassdatum* is date of issue.

Summing up, the preprocessing step focuses on cleaning string and date values that will be the key features for matching the records of the archive and the genealogy.

metrics like Levenshtein distance that are not guaranteed to return a value between 0 and 1 are normalized. These values are computed between the date of the record in the genealogy and each of the date values in the archive: date of issue, entry into force, and year of publication.

String values on the other hand can be compared over edit distances. It is often useful to take different types of distances. The *RecordLinkage* library supports multiple metrics, but for this research, we use: Q-Grams, Jaro-Winkler, Levenshtein, and cosine distance. These distances are computed between the occupation field of the genealogy and each of the occupation, title, and title suffix columns in the archive records. An excerpt of the table that contains the record pairs and the string distances is shown in Table 2.

idx-g	idx-a	jw-o-o	jw-o-t	jw-o-ts	lev-o-o	lev-o-t	lev-o-ts	...
0	225	0.900000	0.518921	0.000000	0.500000	0.195652	0.000000	...
0	226	0.900000	0.570426	0.000000	0.500000	0.222222	0.000000	...
...	...	...	...	...	...	...	...	...

Tab. 2: Excerpt of the computed string distances that will be part of the features for classification. *idx-g* is the index of the record in the genealogy, *idx-a* is the record index of the archive record of this pair. *jw* refers to Jaro-Winkler distance, *lev* to Levenshtein distance. *o-o* is the edit distance between the occupation in the genealogy and the occupation in the archive. *o-t* is the edit distance between the occupation in the genealogy and the title in the archive record. *o-ts* is the edit distance between the occupation in the genealogy and the title suffix in the archive.

For dates, the difference between the compared dates can be computed to check their similarities. In its current implementation, exact matching dates are classified as 1, and dates that are not the same as 0. These features will be used to train the classifier in the next step.

4.4 Classification

The last step of the record matching process is the classification. For this part, common clustering and classification methods can be used to separate the matches from the non-matches. To do so, a subset of the record pairs, consisting of 310 matches and 164 non-matches, was labeled. A random selection of 80% of each matching and non-matching pairs was used as train data, the remaining 20% of each as test data.

Using the labeled data, a SVM and a logistic regression classifier have been trained. As an unsupervised alternative to these two classifiers, also a K-Means classifier has been trained. The prediction results of each of the classifiers are shown in the following section.

5 Experimental Results

The sorted neighbor indexing resulted in a reduction ratio of about 99,78%, meaning that less than 1% of all possible record pairs have been kept in the index. For these records,

SVM, logistic regression, and K-means classifiers were used to make predictions on the test data set. Precision, recall, F-score, accuracy, and specificity have been calculated for each classifier. The results are plotted in Figure 2.

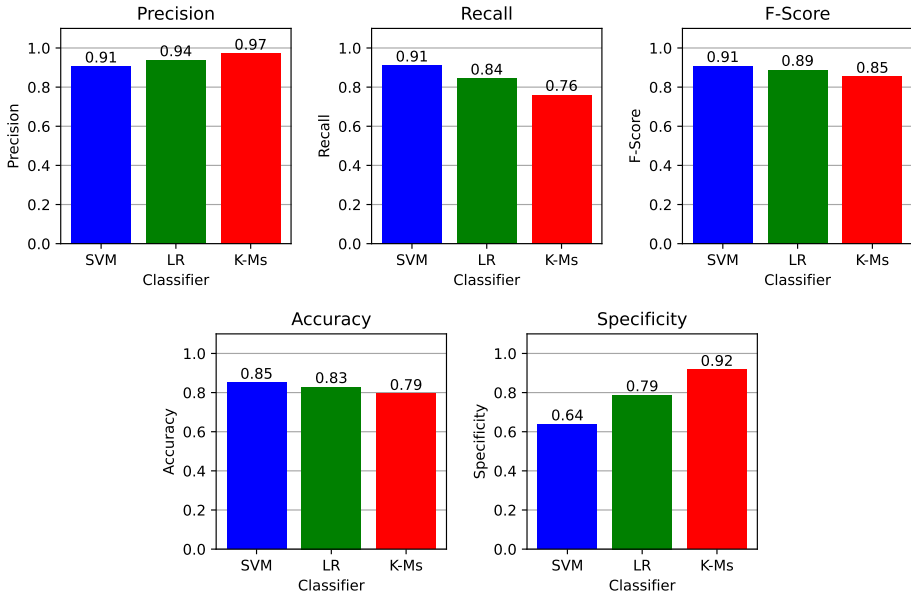


Fig. 2: Performance of SVM, logistic regression and K-Means classifiers for linking records from the genealogy to the archive measured in different metrics.

For the SVM predictions, precision and recall were both at 91%, making it the best classifier when focusing on precision and recall, as seen in the F-score. While it had the lowest value for specificity, showing its lower efficiency for recognizing non-matches, it did not exclude as many true matches as K-Means.

Logistic regression classification lies between SVM and K-Means regarding all considered metrics, making it a useful trade-off between the qualities of the other two classifiers.

The K-Means classifier gives the highest precision with 97%, meaning that it had the highest ratio of correctly predicted true matches. It also had the highest specificity, meaning that it also had the highest fraction of predicted non-matches that were true non-matches. Although it was good in classifying matches and non-matches, it had the lowest recall, so it did find the least true matches in the data set. From all three classifiers, it had the lowest accuracy, so the least overall correct predictions.

Overall, the highest difference in classification quality of all selected classifiers can be found in the recognition of non-matches while precision and accuracy only vary by 6%.

6 Discussion

The reduction ration of more than 99% shows that many potential pairs have been removed right from the start and also reduced the number of non-matches a lot. Additionally, the reduction also impacted the amount of data that becomes relevant for labeling. However, doing so also reduced the number of non-matches in the training set, causing a class imbalance where there are almost twice as many matches as non-matches.

This class imbalance is likely also the reason why K-Means has the lowest accuracy, although it has the highest precision and specificity. The biggest issue with K-Means seems to be due to too many false negatives, which is also why it has the lowest recall.

SVM on the other hand seems to have more trouble recognizing non-matches, which is why it has more false positives, thus decreasing precision, but also improving recall. In cases where correct recognition of matches are more important than gathering as many correct matches and non-matches, K-Means might be a little more helpful. However, for the case of matching records of the genealogy to the archive, this is less critical. While Logistic regression gives a trade-off between all metrics, having more of the actual matches is desirable for the selected data sets.

Another issue with the data might be loading all dates as strings and having the dates in these recognized automatically, which might introduce some errors in cases where the date was not correctly identified in the text. Finally, there could be errors in the data itself which even the record linkage approaches cannot solve. Especially in data sets that have been created by a human, some of the errors might be too severe to fix them in a rule-based manner.

7 Conclusion and Outlook

In this work, we applied record linkage methods on two data sets for historical German VET: The genealogy of vocational education and the occupations archive. While the genealogy data set gives an overview over which occupations and successors of them exist, the occupations archive is supposed to contain a legal regulation for each of the records in the genealogy.

Here, three classifiers that separate record pairs into matches and non-matches have been trained and evaluated: SVM, logistic regression and K-Means. While K-Means was able to be very precise in recognizing matches, there were still many matches that have not been recognized at all. SVM on the other hand was able to find many correct matches, but also allowed for too many false positives and could not recognize non-matches as good as K-Means. The performance of logistic regression lied between the other two classifiers. Additionally, there is still a major class imbalance in the data sets where there are nearly twice as many record pairs that are matches than pairs with non-matches

In future research, more training data and features as well as potentially optimized indexing algorithms could improve the results. Approaches like this can help identify gaps in the occupations archive where certain regulation documents should exist but cannot be found in the archive records. On the other hand, an actively maintained archive could also help identifying errors and gaps in the genealogy, so each of the complementary data sets could be used to refine the other. Additional approaches that remain to be investigated is the training of deep neural networks which could also be trained on the generated features. Large language models might also be subject of future research due to their capability of handling natural language data. As the focus of this research was based on methods that are natively supported by the recordlinkage toolkit [De19], these methods are not part of our presented method.

It also remains to investigate record pairs that have not been classified correctly to further identify challenges in this approach when working on labor market data. Although the classifiers are able to give good results, there are some challenges like title names in the archive that can be very long and only contain a short substring with an occupation that can be mapped to the genealogy. The approach also relies on a field in the archive that stores the occupation name which is quite specific for such a data set. Information like this where some keywords are given in such a structured format might not be applicable for all archives.

References

- [Al24] Alves, A. L. F. et al.: Cross-Lingual Learning Strategies for Improving Product Matching Quality. In: *Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing*. Pp. 313–320, 2024.
- [Ca77] Campbell, D. T. et al.: Confidentiality-preserving modes of access to files and to interfile exchange for useful statistical analysis. *Evaluation Quarterly* 1 (2), pp. 269–300, 1977.
- [Ch12] Christen, P.: *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Springer Publishing Company, Incorporated, 2012.
- [Co12] Cox, S. et al.: The development of a data-matching algorithm to define the ‘case patient’. *Australian Health Review* 37 (1), pp. 54–59, 2012.
- [Da07] Da Silva, R. et al.: Measuring quality of similarity functions in approximate data matching. *Journal of Informetrics* 1 (1), pp. 35–46, 2007.
- [De19] De Bruin, J.: *Python Record Linkage Toolkit: A toolkit for record linkage and duplicate detection in Python, version v0.14*, 2019, DOI: 10.5281/zenodo.3559043.
- [Gu22] Guardiolle, V. et al.: Linking biomedical data warehouse records with the national mortality database in France: large-scale matching algorithm. *JMIR Medical Informatics* 10 (11), e36711, 2022.
- [JSS02] Johannesmeyer, M. C.; Singhal, A.; Seborg, D. E.: Pattern matching in historical data. *AICHe journal* 48 (9), pp. 2022–2038, 2002.
- [Ka75] Kadane, J. B.: *Statistical problems of merged data files*. 1975.
- [LKC06] Lafon, S.; Keller, Y.; Coifman, R. R.: Data fusion and multicue data matching by diffusion maps. *IEEE Transactions on pattern analysis and machine intelligence* 28 (11), pp. 1784–1797, 2006.

- [Ok74] Okner, B.: Data matching and merging: an overview. *Annals of Economic and Social Measurement*, Volume 3, number 2, pp. 347–352, 1974.
- [Ol25] Olar, A.: Resolvi: A Reference Architecture for Extensible, Scalable and Interoperable Entity Resolution. *arXiv preprint arXiv:2503.08087*, 2025.
- [OM08] Olteanu Raimond, A.-M.; Mustière, S.: Data matching—a matter of belief. In: *Headway in Spatial Data Handling: 13th International Symposium on Spatial Data Handling*. Springer, pp. 501–519, 2008.
- [Pa07] Patel, C. et al.: Matching patient records to clinical trials using ontologies. In: *International Semantic Web Conference*. Springer, pp. 816–829, 2007.
- [Sc07] Scannapieco, M. et al.: Privacy preserving schema and data matching. In: *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*. Pp. 653–664, 2007.
- [St05] Steinbock, D. J.: Data matching, data mining, and due process. *Ga. L. Rev.* 40, p. 1, 2005.
- [WF99] Walter, V.; Fritsch, D.: Matching spatial data sets: a statistical approach. *International Journal of geographical information science* 13 (5), pp. 445–473, 1999.