Linking Vocational Archive Data Using an Occupations and Educations Centric Ontology

Thomas Reiser[1,*,†], Jens Dörpinghaus[1,2,3,*,†], Petra Steiner[2] and Michael Tiemann[1,2]

[1]University of Koblenz, Department of Computer Science, Germany
[2]Federal Institute for Vocational Education and Training (BIBB), Bonn, Germany
[3]Linnaeus University, Department of Computer Science and Media Technology, Växjö, Sweden

Abstract

In this paper, an approach is presented for semantically enriching and linking historical vocational education and training (VET) documents using an ontology-centric method grounded in occupations and educational programs. The present study draws on a digitized corpus of archival documents from various political regimes in Germany—including the German Empire, the German Democratic Republic (GDR), and the Federal Republic Germany (FRG) — in order to explore strategies for annotating job titles and aligning them with standardized taxonomies such as KldB and ISCO. The proposed methodology integrates phrase matching, classification models, and ontology-based linking via the German Labor Market Ontology (GLMO), thereby enabling cross-referencing of documents by occupation and educational structure. The proposed workflow is designed to support longitudinal studies and promote interoperability across fragmented archival collections. This offers a scalable solution for labor market and education research.

Keywords

Text analysis, NER, data linking, computational social sciences

1. Introduction

Vocational education and training (VET) systems are of critical importance in maintaining a skilled workforce. In Germany, a historically extensive corpus of VET and continuing VET (CVET) regulations has been published in the Federal Gazette over the course of several decades. These documents serve as the foundational elements of occupational standards and training frameworks. However, the archival form of these regulations as described in [1]—primarily as unstructured or semi-structured scanned documents—poses challenges for digital accessibility, analysis, and integration with contemporary data systems. The digitization of archival material presents an opportunity to preserve, structure, and analyze regulatory knowledge in a form amenable to semantic linking, machine learning, and long-term data curation as discussed in our previous work [2, 3].

Linked data for archival information is not a new, but a crucial topic [4] to guarantee that data can be re-used, is findable and interoperable. This quickly lead to methods developed within the context of the *Semantic Web*, as semantics (the 'meaning' of data) require structure and a technical implementation of describing elements like controlled vocabularies and ontologies. Approaches like the '5 Stars Open Data' initiative (see https://www.w3.org/DesignIssues/LinkedData.html) formulated a popular approach to store open data, including linked data using the RDF standard (Resource Description Framework). Since the open license is often not possible for various reasons (confidential data, proprietary format of a certain device, ...), alternatively the application of the FAIR concept to the data is often more considerable. The FAIR Guiding Principles for scientific data management and stewardship were published in 2016 by Wilkinson *et al.* [5].

Perspectives on Humanities-Centred AI and Formal & Cognitive Reasoning Workshop 2025,(CHAI 2025 & FCR 2025), Joint Workshop at the 48th German Conference on Artificial Intelligence, September 16, 2025, Potsdam, Germany

*Corresponding author.
†These authors contributed equally.
treiser@uni-koblenz.de (T. Reiser); doerpinghaus@uni-koblenz.de (J. Dörpinghaus); steiner@bibb.de (P. Steiner); tiemann@bibb.de (M. Tiemann)
0009-0007-5452-7800 (T. Reiser); 0000-0003-0245-7752 (J. Dörpinghaus); 0000-0001-7136-2744 (M. Tiemann)

To facilitate comprehensive longitudinal analyses of vocational development in Germany, the German Labor Market Ontology (GLMO, see [6, 7]) is being extended with historical occupational taxonomies from both the Federal Republic of Germany (FRG) and the former German Democratic Republic (GDR). This ontological enrichment involves the alignment of legacy classification systems, such as KldB 1988 and KldB 1992, with more recent taxonomies, including KldB 2010 and ISCO-08, through a series of conversion mappings. Beside information about occupations, GLMO also holds data on educational programs, for example vocational education and training (VET) and continuing training (CVET), but also academic programs. All data is mapped to occupations. Thus, the GLMO could act as a linking layer for the variety of labor market data in the archive which we will present in Section 3.1.

In this context, two fundamental questions emerge: First, how can we link a wide variety of documents in an occupational archive based on occupations and educational programs? Second, what are technical and methodological approaches to efficiently annotate and link these documents? For this, we will describe two usecases:

Linking Documents by Title In the pursuit of historical knowledge, the researcher must scrupulously document and analyze related documents from disparate decades. In her research, she encountered references to other texts, including "Berufsbildungsplan für den Lehrberuf Zuckerbäcker" and "Prüfungsanforderungen für Zuckerbäcker," both from the 1930s. These references are not formally cited, but rather, they are implicitly referenced through titles or occupational names.

Semantic linking by document titles facilitates her analysis of the transformation of the Zuckerbäcker role into the modern Konditor/in, supported by archival documents from 1934, 1983, and 2003. Absent this mechanism, references remain isolated, and longitudinal research on occupational evolution becomes significantly more arduous.

Linking Job Titles to KldB Enables Grouping Across Variants Labor market analysts are engaged in the study of the evolution of technical and skilled occupations between the German Democratic Republic (GDR) and the Federal Republic of Germany (FRG). Titles such as Elektromechaniker, encountered in documents from the GDR era, are compared with Elektroniker/in für Betriebstechnik, a title that emerged in post-1990 regulations. A comparison of the two systems reveals that they are mapped to analogous regions in KldB2010, thereby facilitating semantic alignment despite the presence of disparate naming conventions, political contexts, and institutional origins.

Furthermore, the analyst examines the correlation between specific vocational training pathways, such as engineering, and subsequent academic professions, including the Diplom-Ingenieur qualification. KldB-based linking facilitates the examination of transitions from apprenticeship-based careers in the Third Reich or GDR to higher education tracks in the FRG. Absent this general classification layer, identifying structural equivalences and educational mobility across systems and historical regimes would be highly impractical.

Our work is organized into five sections. The initial section delineates the historical context and existing ontologies as well as potential use cases for our contribution. The second section is devoted to a review of the extant literature on linked data and occupational classification. In Section 3, the archive data is presented, along with the occupational and educational taxonomies. It also introduces methods for annotation and linking of labor market data. Then, the fourth section presents the experimental results, while the fifth and final section offers conclusions and recommendations for future research directions.

2. Related Work

Linked data principles have already been adapted in archival contexts in past research, see [4] or in the context of UK National Archives [8]. Initial steps for such an adaption include cleaning, linking and publishing the available data, for which several approaches can successfully be utilized [9, 10]. However, using methods from the semantic web has been found to be most suitable [11, 12]. The automated annotation of occupational data remains a particular challenge.

The integration of diverse labor market data sources is acknowledged as a multifaceted undertaking [13]. However, this work focuses on the mapping and automated classification of job titles in the German language. While dictionary-based methods are commonly employed, machine learning (ML)-based approaches have also been explored. Existing training datasets are often compiled from survey responses or classification systems, including KldB, ESCO, and other synonym collections.

The automated classification of job titles is a significant topic in both academic research and for practitioners in labor market analysis. It is also applicable to many other use cases, including occupations in literature and other texts, such as parliamentary debates [14]. In certain applications, such as surveys, it is necessary to map data to standardized classifications. Two prominent examples of such classifications are the German Classification of Occupations (KldB) and the International Standard Classification of Occupations (ISCO). Additional use cases include the classification of online job advertisements (OJAs) and the alignment of occupational titles from other sources, such as online platforms like Kununu [15].

A multitude of classification categories are recognized for occupations. The International Standard Classification of Occupations (ISCO) was developed by the International Labor Organization (ILO) and published in 1958, 1968, 1988, and most recently in 2008)[1]. The ISCO 2008 has also been utilized within the European Union (EU), with certain German-speaking countries (Germany, Austria, and Switzerland) developing a customized version of the classification. The International Standard Classification of Occupations (ISCO) is structured at a skill level and linked to the "European Skills, Competences, Qualifications and Occupations" (ESCO) ontology, which adds another hierarchy level to the data. In Germany, the Classification of Occupations (KldB) serves as the reference classification for the Federal Employment Agency (BA) and its research institute (IAB)[2]. In this organization, occupations are structured at a task level. The most recent version is the 2020 revision of the KldB 2010, which has undergone a comprehensive redesign, thereby rendering the previous versions from 1988 and 1992 obsolete. The development of this system was undertaken with the objective of ensuring compatibility with the ISCO-08 standard. The study of job titles and taxonomies has a long history, extending even before the advent of computer technology [16].

A portion of the research has focused on the classification of OJAs according to the O*NET framework [17]. This has included the application of normalization approaches [18] and similarity-based methods [19]. The classification of job titles is also employed in the context of online job recruitment [20].

A limited number of publications have been published on the subject of German job titles, with a particular emphasis on the German KldB, for instance, a technical report based on OJAs [21]. While challenges on level 4, which represents a more accurate representation of occupations in the KldB, remain, promising results on level 1 were obtained, which already indicates in which area the occupation is settled. Malte Schierholz's 2018 publication [22] introduced the concept of auxiliary classifications in the field of occupational coding. For further research on the subject of occupational coding in surveys, we refer to [23]. A master's thesis endeavors to predict KldB 5-digit job titles from survey data, thereby highlighting the persistent challenges associated with this endeavor, see [24]. In a similar vein, a scholarly article was published that compared the classification of survey data using BERT and GPT-3, see [25]. However, the absence of a standardized reporting methodology precludes the direct comparison of their results. Nevertheless, they evince analogous challenges to those observed in other studies. Our previous study [26] lends support to this assertion, particularly in terms of the conclusion that large language models (LLMs) are not capable of enhancing the quality of automated classifiers to a significant extent. Consequently, both the classification of occupational areas and, in particular, the level of performance (5th digit) persist as arduous tasks.

[1] See https://www.ilo.org/public/english/bureau/stat/isco/isco08/.

[2] See https://statistik.arbeitsagentur.de/DE/Navigation/Grundlagen/Klassifikationen/Klassifikation-der-Berufe/KldB2010-Fassung2020/KldB2010-Fassung2020-Nav.html.

Table 1
Number of documents (top 10 categories) available in archives before BBiG 1969.Some archives are not yet fully examined.

	Before BBiG 1969	
1	Berufsbilder	2,647
2	Prüfungsanforderungen	838
3	Berufsausbildungsplan	352
4	Berufsbildungsplan	270
5	Ausbildungsrichtlinien	221
6	Berufseignungsanforderungen	193
7	Fachliche Vorschriften zur Regelung des Lehrlingswesens und der Gesellenprüfung im Handwerk	165
8	Berufsausbildung der/des ...	147
9	Fachliche Vorschriften für die Meisterprüfung	142
10	Sammeldokumente	90

Table 2
Number of documents (top 10 categories) available in archives in GDR. Some archives are not yet fully examined.

	GDR	
1	Berufsbilder	914
2	Ausbildungsunterlagen	783
3	Programm für die Fachbildung der Meister	128
4	Qualifikationscharakteristik	101
5	Rahmenausbildungsunterlagen	79
6	Berufs- und Qualifikationscharakteristik	59
7	Ergänzungen zu den Ausbildungsunterlagen	52
8	Ausbildungspläne	25
9	Ausrüstungsnormative	24
10	Ausbildungsordnung	9

3. Data and Methods

In this section, the archive data and the data on job titles and educational programs are presented. Subsequently, we will engage in a discussion concerning methodologies for annotating and linking the archive data. We follow a study of automated classifications presented in [26].

3.1. Archive Data

The available documents within the occupational archive cover a long period of time, but also several states (the German Empire, the German Democratic Republic, and the Federal Republic of Germany). Roughly, these archive materials fall into three large groups: Documents relating to decrees before the introduction of the BBiG in 1969, documents on vocational education and training in the Federal Republic of Germany after BBiG and documents from the German Democratic Republic. This collection of occupation-related documents with legal bases, maintained by the Federal Institute for Vocational Education and Training (BIBB), reflects about 85 years of German VET history, see Table 1 for an overview of the top document categories from the three top domains. In recent years, this collection has been systematically recorded for the first time, resulting in precise knowledge of its contents on the one hand and the state of preservation of the individual documents on the other. This leads to a comprehensive list of data available. For details, see Tables 2 and 3.

In addition to other important metadata like publication year, publisher, etc. the two most relevant information in the available data are the occupational name and the document title. For example,

Table 3
Number of documents (top 10 categories) available in archives in BRD after BBiG. Some archives are not yet fully examined.

	BRD, after BBiG	
1	Ausbildungsordnungen	387
2	Rahmenlehrpläne	366
3	Fortbildungsregelungen der zuständigen Stellen	162
4	Regelungen der zuständigen Stellen für die Berufsausbildung von Menschen mit Behinderungen	31
5	Fortbildungsregelungen des Bundes	15
6	Fortbildungsregelung der Länder	11
7	Umschulungen	3
8	Länderrechtliche Ausbildungen im Gesundheitswesen	2
9	Länderrechtliche Fortbildungen im Gesundheitswesen	2

for 'Achatbohrer', we find two documents from the 1930s: 'Berufsbild des Achatbohrers' and 'Prüfungsanforderungen für Achatbohrer'. While this occupation quickly was included in other crafts, other occupations have a history till today. For example, for 'Zuckerbäcker' we find several documents from the 1930s: Three occupation description from different publishers ('Berufsbild des Zuckerbäckers') but also 'Berufsbildungsplan für den Lehrberuf Zuckerbäcker', 'Berufs-Eignungsanforderungen für den Eintritt in den Lehrberuf Zuckerbäcker' and 'Prüfungsanforderungen für Zuckerbäcker'. The modern occupation name is 'Konditor/in' with the corresponding job number (KldB) 29222. For this occupation, we have (re-)regulations, but also other documents like Berufsbilder, from 1934, 1983 and 2003. In summary, all data is both education and occupation centric.

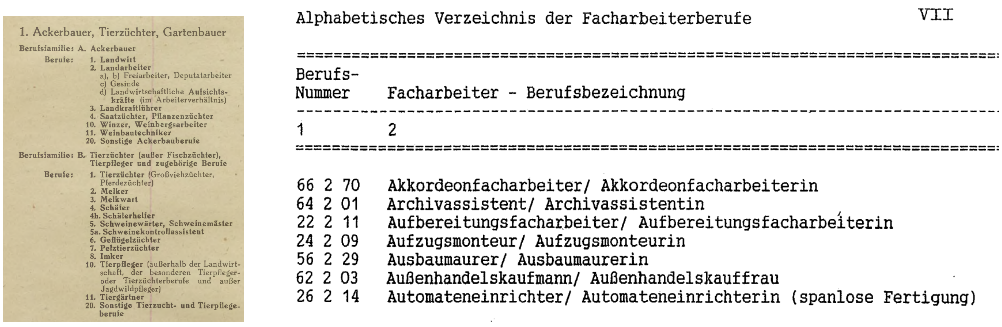
However, while some documents might refer to others by their complete title, for instance, referring to deprecated or existing regulations, these documents are predominantly referred to by occupational names. This assertion is particularly salient in the context of non-legal documents, such as 'Berufsbilder' utilized in the field of job consultation. Therefore, it is insufficient to link the archive by explicit references to other documents by title; implicit references must also be identified.

Currently, this dataset contains digitalized metadata for 2,093 occupations with 7,091 documents (approximately 120,000 pages in total). The documents consist of 4,672 records before BBiG 1969, 1,751 records from GDR, and 614 records from BRD after 1969. The dataset contains different metadata, but not all datasets are linked to a taxonomy of occupations. While this metadata is valuable for researching occupations and their history, an increasing number of documents are available in scanned and digitalized form (TEI-XML). For more information on the digitalized archive and its content, but also methodological approaches we refer to [1, 2, 3, 27, 28]. Thus, the data is already available in a digitalized way and ready for text annotations.

3.2. Data on Job Titles and Educational Programs

For all time periods, we find taxonomies of occupations and education, for example the "Berufsverzeichnis für die Arbeitseinsatzstatistik" from the 1930s, but also from the GDR and mappings between historical and updated taxonomies. For an example of the original data, see Figure 1. As a first step to identify job titles, we can follow the data provided and discussed in [29]. One dataset encompasses 526,535 synonyms and variants of male, female, and neutral job titles. This dataset was provided by the German Federal Employment Agency (BA)[3]. For example, both "Meister – Maßschneiderei" and "Herrenschneidermeisterin" link to KldB 28293. The presence of numerous duplicates results in a non-unique linkage. The most prominent example is the term "Meister" (Master), which is linked to nearly all crafts. Furthermore, the dataset under consideration contains terms associated with occupations. For instance, the terms "Kohle", "Naturwerkstein", and "Anlagenführung" are linked to 21212-129. Consequently,

[3]Available at https://www.arbeitsagentur.de/institutionen/dkz-downloadportal.



1. Ackerbauer, Tierzüchter, Gartenbauer

Berufsfamilie: A. Ackerbauer

Berufe:
1. Landwirt
2. Landarbeiter
   a), b) Freiarbeiter, Deputatarbeiter
   c) Gesinde
   d) Landwirtschaftliche Aufsichtskräfte (im Arbeiterverhältnis)
3. Landkraftführer
4. Saatzüchter, Pflanzenzüchter
10. Winzer, Weinbergsarbeiter
11. Weinbautechniker
20. Sonstige Ackerbauberufe

Berufsfamilie: B. Tierzüchter (außer Fischzüchter), Tierpfleger und zugehörige Berufe

Berufe:
1. Tierzüchter (Großviehzüchter, Pferdezüchter)
2. Melker
3. Melkwart
4. Schäfer
4h. Schäferhelfer
5. Schweinewärter, Schweinemäster
5a. Schweinekontrollassistent
6. Geflügelzüchter
7. Pelztierzüchter
8. Imker
10. Tierpfleger (außerhalb der Landwirtschaft, der besonderen Tierpfleger- oder Tierzüchterberufe und außer Jagdwildpfleger)
11. Tiergärtner
20. Sonstige Tierzucht- und Tierpflegeberufe

Alphabetisches Verzeichnis der Facharbeiterberufe VII

Berufs-Nummer	Facharbeiter - Berufsbezeichnung
1	2
66 2 70	Akkordeonfacharbeiter/ Akkordeonfacharbeiterin
64 2 01	Archivassistent/ Archivassistentin
22 2 11	Aufbereitungsfacharbeiter/ Aufbereitungsfacharbeiterin
24 2 09	Aufzugsmonteur/ Aufzugsmonteurin
56 2 29	Ausbaumaurer/ Ausbaumaurerin
62 2 03	Außenhandelskaufmann/ Außenhandelskauffrau
26 2 14	Automateneinrichter/ Automateneinrichterin (spanlose Fertigung)

Figure 1: Job titles from "Berufsverzeichnis für die Arbeitseinsatzstatistik" (1930s, left) and "Facharbeiterberufe der Deutschen Demokratischen Republik und zugeordnete vergleichbare Ausbildungsberufe der Bundesrepublik Deutschland", (1990, right)

this dataset can also be utilized to identify any implicit relations in a text that link to occupations. Nevertheless, for the aforementioned approach, a blacklist of these terms was created. While these terms are classified under the gender-neutral section, the removal of these terms would also result in the removal of all gender-neutral occupational titles.

This dataset is supplementary information derived from the 'Datensystem Auszubildende' (DAZUBI) database, which is provided by the Federal Institute for Vocational Education and Training (BIBB). The scope of the data encompasses vocational training statistics from Federal and State Statistical Offices, along with annual survey data concerning vocational training for trainees, contracts, and examinations. However, the dataset utilized exclusively contains current and deprecated labels, which are employed for training, retraining, and other vocational education purposes. This modest dataset encompasses 1,875 labels, which frequently contain supplementary information, such as the examination organization. Consequently, the terms 'Änderungsschneider/-in (Hw)' and 'Änderungsschneider/-in (IH)' are encountered.

3.3. Annotating job titles

The process of associating job titles with archive data can be accomplished through a variety of methodologies. Given the availability of a substantial set of word lists, the implementation of phrase matchers is a viable option [26]. However, the German language poses unique challenges in the context of occupation classification. Occupational titles may manifest not only as single nouns but also as complex noun phrases, potentially including information such as certifying institutions (e.g., IHK). Additionally, there are gender-specific and gender-neutral variants of titles. A further complication arises in general texts from surnames that derive from historical professions. A notable example is the surname "Bäcker", which is derived from the German word for "baker". Examples from the Plenary Proceedings of the German Bundestag include:

- Carsten Schneider spricht jetzt für die SPD-Fraktion. (Carsten Schneider now speaks on behalf of the SPD parliamentary group.)
- Da wird Herr Weber nicht begeistert sein! (Mr. Weber will not be thrilled!)
- Stefan Müller [Erlangen] [CDU/CSU]

Such references are not frequently encountered in purely descriptive or even legal documents but particularly evident in author information and references to individuals who can be contacted.

In prior studies, we expanded upon the incorporation of job titles derived from online job advertisements and titles from vocational education and training (VET) to detect KldB entities and exactly match labels and single occupations (see [26]). Additionally, in [29], a module was developed to ascertain whether a recognized term is a family name or a profession.

Similar to this method, for the purpose of annotating job titles in the archive data, a phrase matcher and extensive wordlists are utilized. For this purpose, we operate under the assumption that the digital

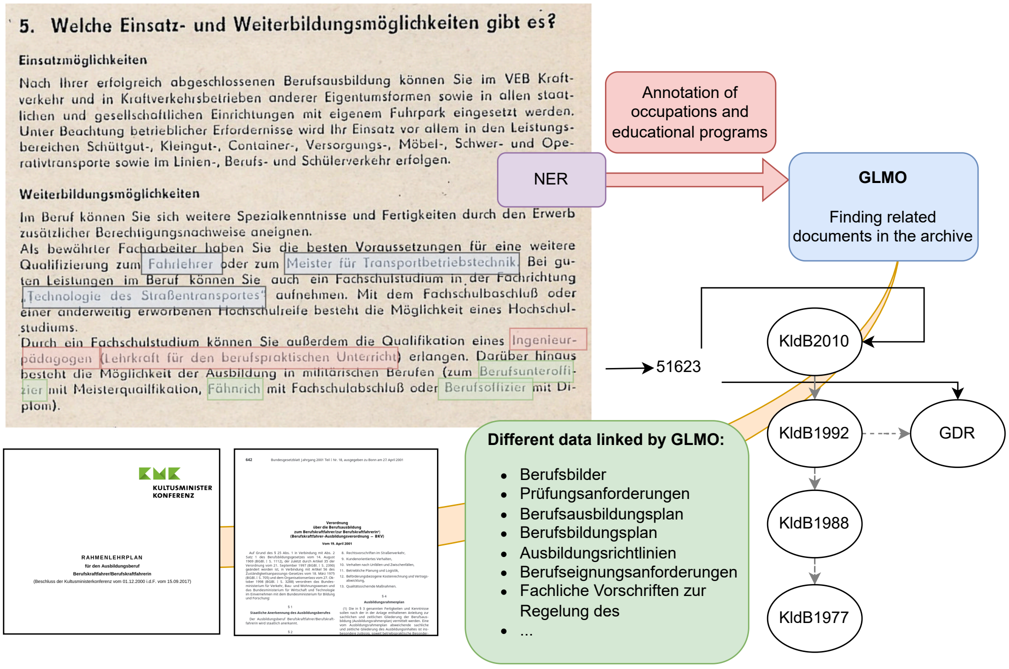


Figure 2: Proposed workflow

data is of a satisfactory quality, devoid of any OCR artifacts. This assumption will likely affect the results because the selected documents were mostly not born-digital. In a subsequent step, a classification model will be implemented to determine the relevance of the reference to an occupation. All data will be linked to KldB2010.

3.4. Proposed Workflow

The proposed workflow consists of three key steps: first, the identification of job titles in archival texts; second, the disambiguation of references to determine whether they denote individuals or occupations; and third, the linking of the results to occupational classifications and related documents. Initial detection is based on extensive synonym lists and phrase matchers, supported by filtering rules to address ambiguities such as surnames derived from professions (e.g., Bäcker). A subsequent classification model is employed to determine whether a detected term refers to an actual occupation or not. The present study utilizes Python 3.11.2 and the Spacy library. For a visual representation of the complete workflow, refer to Figure 2.

Subsequent to the annotation of job titles, they are linked to standardized codes in KldB2010 via historical mappings and synonym datasets. This approach facilitates the establishment of semantic groups across a range of occupational variants, regimes (e.g., GDR vs. FRG), and educational levels. The German Labor Market Ontology (GLMO) facilitates the integration of documents through shared occupational or educational references, enabling the retrieval of related records across temporal and classification systems. To this end, a subset of 30 pages from 17 documents was manually annotated for the purpose of experimental results. The findings are described in the following section.

4. Experimental Results

We employed a variety of records from our data set, including, for instance, Berufsbild documents for "Datenverarbeitungskaufmann" (FRG, 1969), "Datenverarbeitungsfachmann/Datenverarbeitungsfachfrau" (FRG, 1995), Prüfungsanforderungen for "Edelmetallprüfer" (1938), and "Werkgehilfin" (1937).

Table 4
Experimental results for all and non-generic evaluation approaches

Evaluation Approach	Metrics (Macro/Weighted)			
	Precision	Recall	F_1-score	Accuracy
All	**0.45 / 0.82**	**0.50 / 0.91**	**0.48 / 0.86**	**0.91**
Non-generic	0.11 / 0.04	0.50 / 0.21	0.17 / 0.07	0.21
All (unique)	0.39 / 0.60	0.50 / 0.77	0.44 / 0.67	0.77
Non-generic (unique)	**0.26 / 0.27**	**0.50 / 0.52**	**0.34 / 0.36**	**0.52**

Addtionally, legal documents from the Bundesanzeiger for Fahrzeugschlosser (GDR, 1977) were consulted, but also documents for a few other occupations, including "Bekanntmachung zur Verordnung über die Berufsbildung zum Kaufmann für Bürokommunikation/zur Kauffrau für Bürokommunikation nebst Rahmenlehrplan" (1990). Given the presence of two distinct use cases, a dual evaluation strategy was employed, involving the implementation of two distinct evaluation approaches. Either *all* occupations are considered, or exclusively those which are *non-generic*. For instance, the occupation of "Ingenieur für Verfahrenstechnik" is associated with the KldB 25103 and 25104 classifications. However, in this context, the term "Ingenieur" is associated with fourteen additional occupations of a more generic nature. The aforementioned observations also pertain to "Meister", "Lehrkraft", "Kaufmann", and "Kauffrau". Given the prevalence of these terms within the evaluation data, our multifaceted approach ensures a comprehensive analysis. This entails the evaluation of unique data, wherein instances of the same term accompanied by distinct KldB annotations are recorded only once.

The results are presented in Table 4. Issues with all selected occupation detection approaches mainly occur due to OCR problems or the detection of artifacts like "Lohn- und Gehaltsabrechnung." Although the latter can be excluded using a blacklist, problems with the data quality persist as a significant challenge. Nevertheless, with regard to all other issues, the problem lies in detecting too much rather than too little. This renders the approach suitable for existing data, provided that models are re-executed when data quality is enhanced, for instance, when manually corrected. However, when considering the non-generic detection, a different picture emerges. Given that generic occupations generally result in a substantial list of different KldB numbers, we have opted to apply a unique perspective. However, this approach has yielded rather poor results. Consequently, implementing this approach necessitates a substantial investment of manual effort to blacklist generic occupation names or to present these generic occupations to the user.

In the context of this particular use case, it appears more advantageous to enhance the application's front end rather than making refinements to the approach. One potential enhancement would be to accentuate longer annotations.

5. Conclusions and Outlook

The present study proposed a methodology for the detection, annotation, and linkage of occupational and educational references within a substantial corpus of historical vocational documents. The approach integrates rule-based extraction, classification models, and ontology-based linking through the utilization of GLMO. The database utilizes over 500,000 job title variants to identify occupations across time periods, document types, and political regimes. The findings substantiate the viability of high-precision rule-based techniques in structured texts, while underscoring challenges such as data quality issues and ambiguity in generic titles like Meister or Ingenieur.

A significant extension of the approach entailed the direct annotation of document titles, which frequently encapsulate fundamental occupational and regulatory content. Titles such as Prüfungsanforderungen für Zuckerbäcker or Berufsausbildungsplan für Werkzeugmacher contain sufficient semantic information to support classification, even in the absence of document body analysis. Nonetheless, the correlation of such titles with occupations and taxonomies remains an unresolved issue and was not incorporated into our analysis, as not all documents have yet been annotated with KldB or other taxonomy numbers. The long-term perspective for linking the data would be to facilitate efficient retrieval, clustering, and longitudinal comparison of archived regulations, especially in cases where the body text is inaccessible or inconsistent.

In light of these findings, this work suggests several promising areas for future research. Initially, the utilization of alternative artificial intelligence methodologies, such as BERT, as outlined in the work by [14]. Secondly, the integration of rule-based methodologies with LLMs within hybrid systems has the potential to achieve a harmonious equilibrium between precision and adaptability, see for example [30, 31].. The augmentation of training data to encompass a more diverse array of contemporary sources, including user-generated content from online platforms and digitized historical documents, has the potential to enhance the robustness and applicability of models.

Moreover, the implementation of cross-lingual mapping and alignment with international classification systems, such as ISCO and ESCO, would expand the practical relevance of these methods beyond national boundaries. However, future research should also prioritize the development of more robust front-end interfaces to manage generic title ambiguity and facilitate user interaction with linked results.

In summary, the present study provides a substantial dataset and a series of methodological insights that establish the foundation for more precise and extensible systems in occupational text analysis. The findings are of particular pertinence for applications in labor market research, policy analysis, and computational social science, where the automated identification of occupational information persists as a salient concern.

Declaration on Generative AI

During the preparation of this work, the authors used DeepL in order to: Grammar and spelling check. After using these tool(s)/service(s), the authors reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] T. Reiser, J. Dörpinghaus, P. Steiner, M. Tiemann, Towards a datatset of digitalized historical german vet and cvet regulations, Data 9 (2024).

[2] T. Reiser, J. Dörpinghaus, P. Steiner, Analyzing historical legal textcorpora: German vet and cvet regulations, in: INFORMATIK 2024, Gesellschaft für Informatik eV, 2024, pp. 2007–2018.

[3] T. Reiser, J. Dörpinghaus, P. Steiner, Learning from historical vet and cvet regulations in germany: What should vet look like and whom should it serve?, in: NORDYRK 2024 BOOK OF ABSTRACTS, 2025, p. 75.

[4] J. Niu, Linked data for archives, Archivaria 82 (2016) 83–110.

[5] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, et al., The fair guiding principles for scientific data management and stewardship, Scientific data 3 (2016).

[6] J. Dörpinghaus, J. Binnewitt, S. Winnige, K. Hein, K. Krüger, Towards a german labor market ontology: Challenges and applications, Applied Ontology 18 (2023) 343–365.

[7] D. Martić, A. Fischer, J. Dörpinghaus, Extending the german labor market ontology with online data, in: INFORMATIK 2024, Gesellschaft für Informatik eV, 2024, pp. 2019–2030.

[8] P. Clough, J. Tang, M. M. Hall, A. Warner, Linking archival data to location: a case study at the uk national archives, in: Aslib Proceedings, volume 63, Emerald Group Publishing Limited, 2011, pp. 127–147.

[9] S. Van Hooland, R. Verborgh, Linked Data for Libraries, Archives and Museums: How to clean, link and publish your metadata, Facet publishing, 2014.

[10] K. F. Gracy, Archival description and linked data: a preliminary study of opportunities and implementation challenges, Archival Science 15 (2015) 239–294.
[11] B. Wright, O. Brunner, B. Nebel, On the importance of a research data archive, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 32, 2018.
[12] A. Hawkins, Archives, linked data and the digital humanities: increasing access to digitised and born-digital archives via the semantic web, Archival Science 22 (2022) 319–344.
[13] A. Fischer, J. Dörpinghaus, Web mining of online resources for german labor market research and education: Finding the ground truth?, Knowledge 4 (2024) 51–67.
[14] J. Binnewitt, Recognising occupational titles in german parliamentary debates, in: Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024), 2024, pp. 221–230.
[15] K. Hein, J. Dörpinghaus, What is said about vet on social media in germany? trends, demands, and opinions., in: NORDYRK BOOK OF ABSTRACTS, 2024, p. 109.
[16] N. R. Council, D. of Behavioral, S. Sciences, C. on Occupational Classification, Analysis, Work, jobs, and occupations: A critical review of the dictionary of occupational titles (1980).
[17] F. Javed, M. McNair, F. Jacob, M. Zhao, Towards a job title classification system, arXiv preprint arXiv:1606.00917 (2016).
[18] Y. Zhu, F. Javed, O. Ozturk, Document embedding strategies for job title classification., in: FLAIRS, 2017, pp. 221–226.
[19] I. Rahhal, K. M. Carley, I. Kassou, M. Ghogho, Two stage job title identification system for online job advertisements, IEEE Access 11 (2023) 19073–19092.
[20] F. Javed, Q. Luo, M. McNair, F. Jacob, M. Zhao, T. S. Kang, Carotene: A job title classification system for the online recruitment domain, in: 2015 IEEE First International Conference on Big Data Computing Service and Applications, IEEE, 2015, pp. 286–293.
[21] R. Baskaran, J. Müller, Classification of german job titles in online job postings using the kldb-2010 taxonomy (2023).
[22] M. Schierholz, An auxiliary classification with work activity descriptions for occupation coding, AStA Wirtschafts-und Sozialstatistisches Archiv 12 (2018) 285–298.
[23] A. Müller, The implementation of the German Classification of Occupations 2010 in the IAB Job Vacancy Survey: documentation of the implementation process, Technical Report, IAB-Forschungsbericht, 2014.
[24] V. P. V. Karanam, Occupation coding using a pretrained language model by integrating domain knowledge (2022).
[25] P. Safikhani, H. Avetisyan, D. Föste-Eggers, D. Broneske, Automated occupation coding with hierarchical features: A data-centric approach to classification with pre-trained language models, Discover Artificial Intelligence 3 (2023) 6.
[26] R. Dorau, K. Hein, Towards the automated classification of german job titles according to kldb, in: 205th Conference on Computer Science and Information Systems (FedCSIS), 2025.
[27] K. Hein, Linked labor market data: Towards a novel data housing strategy, in: Proceedings of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS), 2024, pp. 355–362.
[28] T. Reiser, J. Dörpinghaus, P. Steiner, Analyzing historical german vet textcorpora: A novel information system, in: Nordyrk Conference 2025 Book of abstracts, 2025, pp. 126–127.
[29] T. Reiser, J. Dörpinghaus, M. Tiemann, Detecting occupations in german texts: Challenges and data, in: Proceedings of the 2nd International Workshop on AI in Society, Education and Educational Research (AISEER), 2025.
[30] S. Laqrichi, A hybrid framework for cosmic measurement: Combining large language models with a rule-based system, IWSM-Mensura (2024).
[31] M. Billi, A. Parenti, G. Pisano, M. Sanchi, A hybrid approach for accessible rule-based reasoning through large language models, in: 18th International Workshop on Juris-Informatics, 2024.