

Towards a Dataset of Digitalized Historical German VET and CVET Regulations

Thomas Reiser ^{1,†}, Jens Dörpinghaus ^{1,2,3,*} , Petra Steiner ²  and Michael Tiemann ^{1,2} 

¹ Department of Computer Science, University of Koblenz, 56070 Koblenz, Germany; reiser@uni-koblenz.de (T.R.); tiemann@bibb.de (M.T.)

² Federal Institute for Vocational Education and Training (BIBB), 53113 Bonn, Germany; steiner@bibb.de

³ Department of Computer Science and Media Technology, Linnaeus University, 352 52 Växjö, Sweden

* Correspondence: doerpinghaus@uni-koblenz.de or jens.doerpinghaus@bibb.de

† These authors contributed equally to this work.

Abstract: The digitization of historical documents has gained particular interest in recent years in the digital humanities. The goal is to digitize historical documents by extracting and structuring text from scanned images. Here, we focus on the processing of historical German VET (vocational education and training) and CVET (continuing vocational education and training) regulations to support educational research. This dataset contains data from 1908 to the present and includes 2125 documents as PDF, 983 fully converted XML documents, and additional metadata for 7090 documents from the archive. We present an overview of the historical background and the challenges of processing different historical documents from three different federal states.

Dataset: <https://doi.org/10.5281/zenodo.10810060>.

Dataset License: CC BY-NC-ND 4.0.

Keywords: digitization; labor market research; TEI-XML; OCR; legal texts



Citation: Reiser, T.; Dörpinghaus, J.; Steiner, P.; Tiemann, M. Towards a Dataset of Digitalized Historical German VET and CVET Regulations. *Data* **2024**, *9*, 128. <https://doi.org/10.3390/data9110128>

Academic Editor: Antonio Sarasa-Cabezuelo

Received: 3 August 2024

Revised: 7 October 2024

Accepted: 16 October 2024

Published: 3 November 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

For tertiary education, the German education system offers different pathways for professionals [1] and it is necessary to distinguish between initial vocational education (training, “Ausbildung” or retraining, “Umschulung”) and continuing vocational education, which includes advanced training (“Weiterbildung”, unregulated, e.g., continuing professional development) and upgrading training (“Fortbildung”).

At the beginning of the 20th century, the distinction between unskilled and skilled labor was still unclear, and the designation of skilled occupations was not clearly defined [2]. The first work on classification began as early as 1908, but it did not take practical effect until the 1920s. After 1933, the National Socialist rulers created the political framework to shape occupational regulation work in their own sense [2]. They form one of the earliest complete source basis for vocational education.

In the Federal Republic of Germany, the Vocational Training Act (BBiG) of 1969, which was reformed in 2005, was passed for this purpose [3]. All those involved in vocational education and training participate in the planning and preparation of new occupations or occupations that need to be modernized are (a) the companies and the chambers (employers), (b) the trade unions (employees), (c) the states, and (d) the federal government. Finally, the federal government provides the legal framework for vocational training through laws and regulations. The Federal Institute for Vocational Education and Training (BIBB), founded in 1970 on the basis of the Vocational Training Act (BBiG), prepares the content of the training regulations.

Vocational education is also embedded in a system of lifelong learning of CVET programs. This advanced training is mostly regulated ¹, while upgrading training is comparatively little regulated. We find 1004 (re-)trainings that are regulated by enterprises or by the Crafts and Trade Code, 549 of them are regulated by the German state. The number of informal trainings is much higher.

Understanding the differences between formal and less formal CVET is therefore crucial, since the government can only directly influence the former. Important actors in continuing vocational training in Germany are the following: (a) educational institutions, (b) companies and enterprises, (c) employees, and (d) sponsors. They all have to deal with changing conditions and requirements as a result of transformation processes, such as increasing digitalization [6]. Another key challenge is demographic change, one of the factors leading to considerable labor-shortcomings in the foreseeable future [7]. While it is discussed how better information about chances and challenges in the world of work and for the way to ones (first) occupational position, propositions ranging from occupational orientation in general schools and before up to closer consultation of individuals (for an overview see [8]) continuing vocational education and training can also play a central role. The upgrading or retraining of skills in the workforce might not create new labor, but it might offer opportunities for people whose tasks have been altered by the introduction of automated tools and machines. Enhancing our understanding of continuous training contents and their skills will help determining which trainings to foster and which to reshape.

However, these challenges have changed over time and a key point is to provide a detailed overview and analysis of how requirements, changes, and challenges are reflected in regulatory documents. For example, an important task is to determine what educational content is increasingly being offered and demanded in order to draw conclusions about the development needs of the vocational education system and, in particular, the continuing education system. The research-based development of the vocational education system should not only ensure the competitiveness of the economy at the systemic level, but also help to counteract unemployment and stabilize the social security system, see [9].

However, the historical regulations (see Section 3 for details) are currently not available in a digitized format. Since the available documents cover not only a long period of time, but also several states (the German Empire, the German Democratic Republic, and the Federal Republic of Germany), the challenges for OCR and data infrastructure are manifold. The dataset contains the outcomes of (a) making the BIBB archive (Berufearchiv) accessible, (b) scanning the documents in it, (c) collecting available online documents, and (d) digitizing these documents. We have made as much data as possible available, although some printed material is still under copyright. For scientific purposes, all data are available upon request.

The rest of this paper is organized as follows: We will first provide a brief literature review, describe the available data, and then outline the methodological observations and preliminary results. Our conclusions describing the challenges of the available documents and an outlook on the impact of our research are drawn in the last section.

2. Literature Review

A lot of different research has been conducted on the digitization of documents in recent years. For example, historical Finnish newspapers, see [10], or historical publications of the Bundesanzeiger, see [11]. These approaches are often combined with the extraction of names, city, and other information. After data extraction, the integration of the results as linked open data is often mentioned. There has also been some research on how to model the structure of legal texts. One approach presented by [12] defines the structure of a legal text in Austria by finding sentences in the text and applying NLP methods. Most authors use tesseract as their OCR engine because it is open source, non-commercial, and provides good results.

Other often-used tools are OCR4all, which provides a semi-automatic workflow especially for digitizing historical documents as a web application [13], OCRopus, which is

also open source but does not perform as well as tesseract, and ABBYY finereader, which is a commercial tool and often gives only slightly better results than tesseract, see [14,15].

The historical development of vocational training regulations has only been studied to a very limited extent [2], while the general history and development of the labor market in relation to occupations receives much attention, see [16–19]. Other works focus on the current development of regulations, see [3], and their analysis is also widely considered [20–22]. However, it remains unclear whether this is due to the fact that historical resources are currently not publicly available.

The historical international standard classification of occupations (ISCO) is also a publicly available dataset of comparable occupations that would be a prerequisite for making historical occupations and regulations interoperable. It was introduced in 2002 [23] and is available as a database at [24], where several datasets can be downloaded. However, the list of German occupations in particular is incomplete, since the data do not include references to GDR occupations or information on training. Another relevant dataset is prepared as Ontologie historischer, deutschsprachiger Berufs- und Amtsbezeichnungen (see [25]), but is currently not publicly available. Classifications for GDR occupations are also not yet digitally available, while their mapping to standards like KldB is widely discussed [26,27]. Another dataset is offered as “Genealogie der Berufe”, but is only available as a web service (see [28]). Also worth mentioning is the seminal work by Wolf-Dieter Gewande, who in 1999 for the first time compiled unpublished recognition data and traced the development of more than 1300 occupations to the present, see [29].

3. Data Description

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. The Federal Institute for Vocational Education and Training (BIBB) maintains a collection of occupation-related documents with legal bases, which reflect 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 Table 2 and [30].

Currently, this dataset contains digitalized metadata for 2093 occupations with 7091 documents (approximately 120,000 pages in total). The documents consist of 4672 records before BBiG 1969, 1751 records from GDR, and 614 records from BRD after 1969. While these metadata are valuable for researching occupations and their history, an increasing number of documents are available in scanned and digitalized form (TEI-XML), which we will discuss in the next subsections.

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

	Before BBiG 1969	#
1	Berufsbilder	2647
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

Table 1. *Cont.*

9	Fachliche Vorschriften für die Meisterprüfung	142
10	Sammeldokumente	90
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
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

Table 2. Metadata available for this record. Not all data are always available, e.g., not all data have been scanned or converted to XML [30].

Type	Data Field	Data Type
Document metadata	Title	text
	Subtitle	text
	Published at	date
	Edict date	date
	Taking effect	date
	Authors	text
	Publisher	text
	Location	text
	Organization	text
	Pages	int-int
	Length (pages)	int
	Type	[VET/CVET/...]
	Link to archive number	text
	Link to PDF	text
Regulation	Occupation	text
	Occupation_id	reference to GLMO
Fulltext	Reference to XML file	text

3.1. Materials on Decrees Prior to BBiG 1969

Starting with the foundation of the German Committee for Technical Education (DATSCH) in 1908, documents were created for the standardization of occupations. Different types of texts together form a whole to regulate an occupation. These are job description (“Berufsbild”), examination requirements (“Prüfungsanforderungen”), vocational training plan (“Berufsbildungsplan”), course (“Lehrgang”), professional suitability requirements (“Berufseignungsanforderungen”) and syllabus (“Lehrplan”).

The paper of most of the documents is browned due to its age. About two-thirds of the documents from this period are in Fraktur script; otherwise, Latin script was used in various fonts. While some of the documents are very well preserved, others are badly damaged: there is water damage, glued on notes, perforations, or mold. Some documents have been inscribed or obviously crossed out in later years, especially concerning passages reflecting Nazi ideology (Figure 1, left).

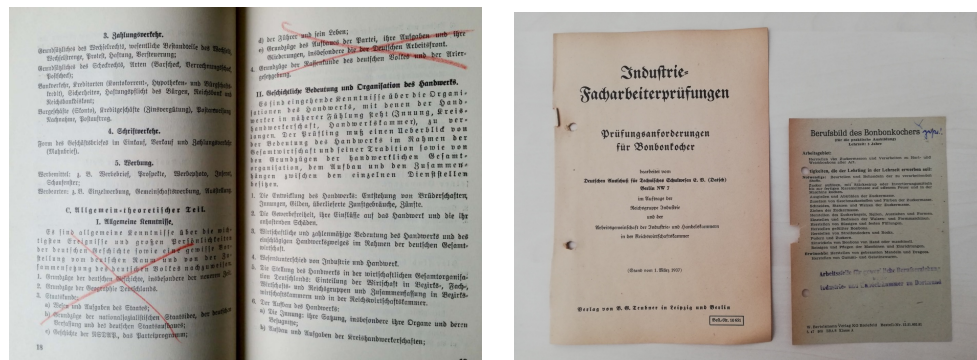


Figure 1. Left: Deutscher Handwerks- und Gewerbeamtstag: Fachliche Vorschriften für die Meisterprüfung im Brunnenbauer-Handwerk. Berlin 1937 (German Chamber of Crafts and Commerce Association: Technical regulations for the master craftsman examination in the well construction trade.). Right: Bonbonkocher (confectioner) 1937, examination requirements and pocket-sized job description.

Most of these early order specifications are in DIN A5 format. However, the range of special formats extends from pocket-sized job descriptions (DIN A6, Figure 1, right) to inserts or glued-in sheets in special formats up to about DIN A1.

3.2. Documents from the Federal Republic of Germany

A smaller part of the collection consists of legal regulations from the Federal Republic of Germany. These can be subject to the BBiG, HwO, specific health profession laws or also the federal school legislation of the German federal states. Accordingly, training regulations, amendment regulations, corrections, framework curricula, and advanced training regulations of the federal government, the federal states, and the competent bodies can be found here.

Some of these documents are already available digitally ². The special challenges for OCR evaluation here are text arranged in two columns and inserted tables.

3.3. Documents of the German Democratic Republic

The available occupation-related materials from the GDR also relate to training and advanced education. The holdings essentially comprise training documents for skilled worker training or training documents for socialist vocational training (“Ausbildungsunterlagen für die Facharbeiteraus- und Weiterbildung” or “Ausbildungsunterlagen für die sozialistische Berufsbildung”), training plans (“Ausbildungspläne”), equipment normatives (“Ausrüstungsnormative”), occupational and qualification characteristics (“Berufs- und Qualifikationscharakteristiken”) as well as various versions of job descriptions (“Berufsbilder”) and programs for the specialized training of master craftsmen (“Programme für die Fachbildung der Meister”).

The GDR materials are more extensive and are often bound as a booklet or book. They contain up to 323 pages and have the normal A4 and A5 formats. However, they are often printed in two columns and in typewriter font (Figure 2). Moreover, the job descriptions for vocational guidance are illustrated with black and white photos and contain color elements (Figure 3).

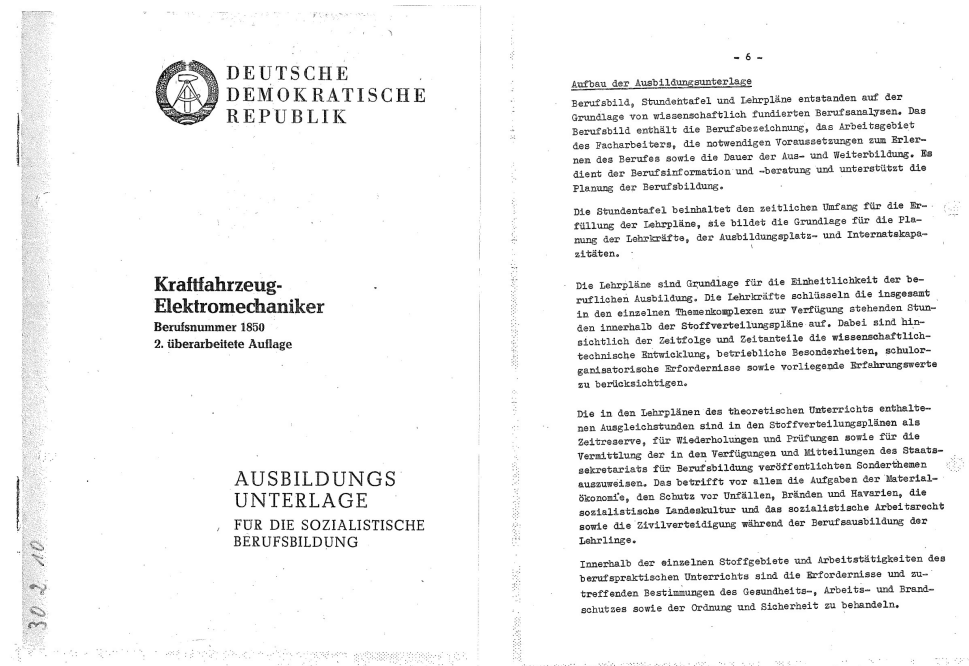


Figure 2. Regulation documents of the German Democratic Republic for electrical mechanical engineers for vehicles (Kraftfahrzeug-Elektromechaniker): cover page (left) and a page describing the curriculum (right). Documents are usually written with typewriters.



Figure 3. Example of job description of vocational guidance for glaziers (Glaser). Title page.

3.4. Dataset Structure

As discussed above, we plan to make all documents available in digitized form. Currently, all data mentioned in Table 1 are available in the archive [30]. However, only a subset is already scanned. Here, we provide a list of available data and the corresponding files. In addition to the scanned documents provided in this dataset, we will also provide XML representations. This dataset is smaller, since we rely on automated pipelines (see next section for details) and manual curation.

Here, all information is encoded in XML according to TEI (Text Encoding Initiative) standards, see [33]. This is an XML standard for organizing text that includes metadata and

text structure, see Figure 4. In the following section, we will briefly summarize the process to create XML files from scanned documents.

§ 2

Ausbildungsdauer, Fachrichtungen

Die Berufsausbildung dauert drei Jahre. Im dritten Ausbildungsjahr kann zwischen den Fachrichtungen

1. Sägeindustrie,
2. Hobelindustrie,
3. Holzwerkstoffindustrie und
4. Holzleimbauindustrie

gewählt werden.

```

1 <div n="1" type="subsection">
2   <head>§ 2 Ausbildungsdauer , Fachrichtungen</head>
3   <div type="subsubsection">
4     <p>Die Berufsausbildung dauert drei Jahre. Im
        dritten Ausbildungsjahr kann zwischen den
        Fachrichtungen</p>
5     <list>
6       <item n="1">
7         <p>1. Sägeindustrie ,</p>
8       </item>
9       <item n="2">
10        <p>2. Hobelindustrie ,</p>
11      </item>
12      <item n="3">
13        <p>3. Holzwerkstoffindustrie und</p>
14      <item n="4">
15        <p>4. Holzleimbauindustrie gewählt werden.</p>
16      </item>

```

Figure 4. Example of TEI XML structure (**bottom**), showing a list of specializations mentioned in a regulation (**top**).

4. Methods

4.1. Data Collection and Archive

The archive material is under scientific research for the last couple of years [34]. Other data are available from federal archives or the Federal Gazette (Bundesanzeiger) which can be found on the website of the archive of the Federal Gazette³ where the VET and CVET regulations have been published since enactment of the BBiG. Crawling the website of the Federal Gazette is not really possible due to its structure⁴. However, there is another website, Offenegesetze⁵, by the Open Knowledge Foundation Deutschland e.V. (OKFDE) that publishes different historical releases and also has an API that is specifically designed to filter for specific documents. While the Federal Gazette has been privatized and only offers an API with a paid subscription, OKFDE offers the data for free. Although many regulations that can be considered are offered, some of the regulations could not be collected due to damaged files or API errors.

However, these different data sources are manually curated by domain experts and offer a comprehensive list of available material. For details, we refer to [36,37]. We will now briefly discuss the pipeline converting scanned documents into TEI XML. For details, we refer to [37,38].

4.2. Pipeline for TEI XML

The pipeline takes multiple PDF and/or image files containing scanned images of historical VET and CVET regulations as input and digitizes them into a processable text format such as hOCR. The pipeline is highly configurable and all steps can be performed with different tools.

Since most scans are skewed to at least some degree and contain noise due to the age of the document, the extracted image files are preprocessed to deskew each page, remove background noise, and binarize the image to contain only black and white pixels, with only the text to be recognized in black. Again, there are already some tools for this task, such as *scantailor*⁶ and *unpaper*⁷. Both can be used via the command line interface and the pipeline can be configured accordingly.

Once the scans are deskewed, denoised, and binarized, the image files are passed to the tesseract OCR engine to create hOCR files that contain the text recognized in the images, as well as layout information about the text. This information is used to structure the data contained in that text. In addition to the generated PDF files that now contain text, the text is converted to a TEI-XML file that further structures the text and adds some meta information such as title, year of publication, converts lists to enumerations, and allows headings to be defined in the text. Figure 2 shows an example output.

While not all records contain PDF scans, all documents with TEI-XML have a reference to a PDF scan as the ground truth. The pipeline recognizes the structure of *Bundesanzeiger* documents and also provides table recognition features. The pipeline produces preliminary results for early documents (before 1945) and for GDR documents, which are not yet included in the dataset. In general, the pipeline provides good results for *Bundesanzeiger* documents, but manual curation and quality control of all data is still an ongoing task.

5. Conclusions and Outlook

In this paper, we introduce the first dataset of digitalized historical German VET and CVET regulations. This dataset contains 983 fully digitized regulations in TEI XML format and 2125 partly digitized regulations as well as metadata for 7090 documents in the *Berufearchiv* covering all regulations which are currently only available as scanned PDF files. The accurate manual annotations available in this dataset should enable researchers to further analyze these textual corpora and help with the design and evaluation of longitudinal labor market data. Access to large textual corpora remains a huge challenge for the social sciences and labor market research. In this context, this dataset is the first contribution to publicly available and manually annotated data.

In this paper, we also described the historical background, the data collection and the challenges of processing historical German VET and CVET regulations using OCR and an approach towards a semi-automatic pipeline for those documents published in the *Bundesanzeiger*. We have shown that not only is there a huge amount of historical documents, but they also vary in format, condition, fonts and formatting. Thus, providing specialized pipelines for different time periods and countries to deliver fully digitized documents is still an ongoing task.

While scant research has been conducted on the historical regulation of vocational education and training in Germany, another research gap emerges: the integration of datasets for occupational classifications not yet available should be accompanied by linked data integration. Thus, the integration of historical datasets such as the KldB 1975, 1988, and 1992 is a crucial step forward. Furthermore, the integration of the data into existing taxonomies (KldB or GDR occupations) remains an open question. Our objective is to provide all data in the German Labor Market Ontology (GLMO). This can be employed for further analysis and enrichment; for example, through the utilization of semantic embeddings [41,42] or integration in existing labor market approaches [43–45], document clustering [46,47], or related social network analysis [48]. However, as demonstrated by [49,50], there are still challenges for automated mapping approaches. At present, our objective is to facilitate both digitization and research through the implementation of a

web application that will enable the tagging, annotation, and search of documents [51]. By making these data accessible to the public and to researchers, we aim to advance the frontiers of labor market research.

Author Contributions: Conceptualization, J.D., P.S. and T.R.; methodology, J.D. and T.R.; software, T.R.; validation, J.D. and P.S.; formal analysis, J.D. and P.S.; investigation, J.D. and T.R.; resources, P.S.; data curation, P.S.; writing—original draft preparation, J.D., P.S. and T.R.; writing—review and editing, J.D., P.S. and T.R.; visualization, T.R.; supervision, J.D.; project administration, J.D. and P.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding. This article was funded by the Open Access Publication Fund of the Federal Institute for Vocational Education and Training (BIBB), Bonn.

Data Availability Statement: All data available at <https://doi.org/10.5281/zenodo.10810060> according to CC BY-NC-ND 4.0 license.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

VET	Vocational Education and Training
CVET	Continuing Vocational Education and Training
DATSCH	German Committee for Technical Education (Deutscher Ausschuss für Technisches Schulwesen)
DKZ	Documentation number, “Dokumentationskennziffer”
MdVWivWi	Mitteilungsblatt der Verwaltung für Wirtschaft im vereinigten Wirtschaftsgebiet
MinWI	Ministerialblatt des Bundesministers für Wirtschaft
BA	German Federal Employment Agency (Bundesagentur für Arbeit)
BIBB	Federal Institute for Vocational Education and Training
KldB	German Classification of Occupations, “Klassifikation der Berufe”

Notes

- ¹ For example at federal level [BBiG/HwO] or by the federal states [4,5]. This sort of accreditation can also be found in other countries, and it allows a quality assurance leading to official recognition and approval by the relevant legislative or profession authorities.
- ² Since 1 January 2023, the official promulgation of federal laws and ordinances in Bundesgesetzblatt has been exclusively on the internet at [31]. The laws and ordinances published in the official paper edition of the Bundesgesetzblatt from 1949–2022 have since been available for download in an online archive (Bundesgesetzblatt (BGBl.)—Verkündungsblatt der Bundesrepublik Deutschland. Online-Archiv der von 1949 bis 2022 erschienenen Ausgaben [32]
- ³ See [32].
- ⁴ The website structure of [32] is mostly written in JavaScript components that hide a lot of the underlying structure and make it harder to scrape the website. An API is only available behind a paid subscription.
- ⁵ See [35].
- ⁶ See [39].
- ⁷ See [40].

References

1. Graf, L.; Lohse, A.P. Advanced skill formation between vocationalization and academization: The governance of professional schools and dual study programmes in Germany. In *Governance Revisited—Challenges and Opportunities for Vocational Education and Training*; Peter Lang: Frankfurt, Germany, 2021.
2. Herkner, V. Grundzüge der Genese und Entwicklung einer korporatistischen Ordnung von Ausbildungsberufen. *Berufsbild. Wiss. Und Prax.-BWP* **2013**, *42*, 16–19.
3. Kuppe, A.M.; Lorig, B.; Schwarz, H.; Stöhr, A. *Ausbildungsordnungen und Wie Sie Entstehen*; Bundesinstitut für Berufsbildung: Bonn, Germany, 2015.
4. Dobischat, R.; Düsseldorf, K.; Dikau, J. Rechtliche und organisatorische Bedingungen der beruflichen Weiterbildung. *Handb. Berufsbild.* **1995**, 427–440. https://doi.org/10.1007/978-3-531-90622-5_33 (accessed on 7 October 2024)

5. Bauer, R.; Bauer, R. Die Debatte über die Zukunft der dualen Berufsausbildung. In *Verberuflichung von Weiterbildung und die Zukunft der dualen Berufsausbildung*. Forschung SoziologieVS Verlag für Sozialwissenschaften, Wiesbaden: Germany, 2000; pp. 21–84.
6. Helmrich, R.; Tiemann, M.; Troltsch, K.; Lukowski, F.; Neuber-Pohl, C.; Lewalder, A.C.; Gunturk-Kuhl, B. *Digitalisierung der Arbeitslandschaften: Keine Polarisierung der Arbeitswelt, aber beschleunigter Strukturwandel und Arbeitsplatzwechsel*; Number 180; Wissenschaftliche Diskussionspapiere; Martin-Luther-Universität Halle-Wittenberg: Halle, Germany, 2016.
7. Maier, T. Es wird knapp: Ergebnisse der siebten Welle der BIBB-IAB-Qualifikations- und Berufsprojektionen bis zum Jahr 2040. In *BIBB Report: Forschungs- und Arbeitsergebnisse aus dem Bundesinstitut für Berufsbildung*; Online-Ressource (20 Seiten); Deutsche Nationalbibliothek: Frankfurt am Main, Germany, 2022; p. 1.
8. Maier, T. Bildungspolitik gegen Fachkräfteengpässe. *Aus Politik und Zeitgeschichte* **2024**, *74*, 39–46.
9. Dobischat, R.; Käßlinger, B.; Molzberger, G.; Münk, D. *Bildung 2.1 für Arbeit 4.0?*; Springer: Berlin/Heidelberg, Germany, 2019.
10. Koistinen, M.; Kettunen, K.; Kervinen, J. How to improve optical character recognition of historical Finnish newspapers using open source Tesseract OCR engine. In *Proceedings of the 8th Language and Technology Conference, LTC 2017*, Poznan, Poland, 17–19 November 2017; pp. 279–283.
11. Hamann, H. The German federal courts dataset 1950–2019: From paper archives to linked open data. *J. Empir. Leg. Stud.* **2019**, *16*, 671–688.
12. Nabizai, A.; Fill, H.G. *Eine Modellierungsmethode zur Visualisierung und Analyse von Gesetzestexten*; Weblaw: Bern, Switzerland, 2017.
13. Reul, C.; Christ, D.; Hartelt, A.; Balbach, N.; Wehner, M.; Springmann, U.; Wick, C.; Grundig, C.; Büttner, A.; Puppe, F. OCR4all—An open-source tool providing a (semi-) automatic OCR workflow for historical printings. *Appl. Sci.* **2019**, *9*, 4853.
14. Heliński, M.; Kmieciak, M.; Parkoła, T. *Report on the Comparison of Tesseract and ABBYY FineReader OCR Engines*; PCSS: Munich, Germany, 2012.
15. Clausner, C.; Antonacopoulos, A.; Pletschacher, S. Efficient and effective OCR engine training. *Int. J. Doc. Anal. Recognit.* **2020**, *23*, 73–88.
16. Wolf, S. Past meets Present—the history of the German Vocational education and training model as a reflection frame to the prospect of the Egyptian model. *Soc. Dimens. Particip. Vocat. Educ. Train.*; **2017**, *5*, 89.
17. Harney, K. Entstehung und Transformation der beruflichen Bildung als Institution—Systemischer Rück- und Ausblick. *Bild. Erzieh.* **2020**, *73*, 346–357.
18. Protsch, P. *Zugang zu Ausbildung: Eine Historisch Vergleichende Perspektive auf den Segmentierten Ausbildungsmarkt in (West-) Deutschland*; Technical Report; WZB Discussion Paper; WZB: Berlin, Germany, 2011.
19. Maier, T. *Die Anwendbarkeit des Erlernten in den Wandelnden Bildungs- und Arbeitslandschaften der 1970er-bis 2000er-Jahre*; Verlag Barbara Budrich: Leverkusen, Germany, 2021.
20. Gessler, M.; Howe, F. From the reality of work to grounded work-based learning in German vocational education and training: Background, concept and tools. *Int. J. Res. Vocat. Educ. Train.* **2015**, *2*, 214–238.
21. Oliver, D. Complexity in vocational education and training governance. *Res. Comp. Int. Educ.* **2010**, *5*, 261–273.
22. Bliem, W.; Petanovitsch, A.; Schmid, K. *Success Factors for the Dual VET System*; Wien: ibw-Forschungsbericht: Wien, Germany, 2015.
23. Leeuwen, M.v.; Maas, I.; Miles, A. *HISCO: Historical International Standard Classification of Occupations*; Leuven University Press: Leuven, Belgium, 2002.
24. Standardized Occupations. Available online: <https://iisg.amsterdam/en/hsndb/standardized-occupations> (accessed on 7 October 2024)
25. Ontologie Historischer, Deutschsprachiger Berufs- und Amtsbezeichnungen. Available online: <https://www.geschichte.uni-halle.de/struktur/hist-data/ontologie/> (accessed on 7 October 2024)
26. Geis, A.J.; Hoffmeyer-Zlotnik, J.H. *Zur Vercodung von Beruf, Branche und Prestige für die DDR*; Campus Verlag: Frankfurt, Germany, 1991; Volume 5.
27. *Klassifikation der Berufe, K. Band 1: Systematischer und Alphabetischer Teil Mit Erläuterungen*; Bundesagentur für Arbeit: Nuremberg, Germany, 2010.
28. Informationen zu Aus- und Fortbildungsberufen. Available online: https://www.bibb.de/dienst/berufesuche/de/index_berufesuche.php/ (accessed on 7 October 2024)
29. Gewande, W.D. *Historische Entwicklung der Staatlich Anerkannten Ausbildungsberufe und Ihrer Ordnungsmittel von 1934–1999: Unter Berücksichtigung der Mit Deutschen Ausbildungsberufen Gleichgestellten Österreichischen Lehrberufe und Gleichwertigen Facharbeiterberufen aus der Ehemaligen DDR*; Zentralamt der Bundesanst. für Arbeit, Geschäftsstelle für Veröff: Nuremberg, Germany, 1999.
30. Steiner, P.; Waechter, M.; Dörpinghaus, J. BIBB Berufearchiv. 2024. Available online: <https://zenodo.org/records/10810060> (accessed on 7 October 2024).
31. Bundesgesetzblatt. Available online: <https://www.recht.bund.de/> (accessed on 7 October 2024)
32. Bundesgesetzblatt BGBl. Online-Archiv 1949 - 2022. Available online: <https://www.bgbl.de/xaver/bgbl/> (accessed on 7 October 2024)
33. Nellhaus, T. XML, TEI, and Digital Libraries in the Humanities. *Portal Libr. Acad.* **2001**, *1*, 257–277.
34. Menne-Haritz, A. Erschließung. In *Handbuch Archiv: Geschichte, Aufgaben, Perspektiven*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 207–217.

35. OffeneGesetze.de – Freier Zugang zu unseren Gesetzen. Available online: <https://offenegesetze.de/> (accessed on 7 October 2024)
36. Fischer, E. *Genealogie der Ausbildungsberufe: Zur Entwicklung der Ausbildungsberufe in Deutschland von 1926–1990*; Bundesinst. für Berufsbildung: Bonn, Germany, 1990.
37. Reiser, T.; Dörpinghaus, J.; Steiner, P. Learning from historical VET and CVET regulations in Germany: What should VET look like and whom should it serve? In Proceedings of the NordYrk Conference 2024, Reykjavik, Iceland, 3–5 June 2024.
38. Udelhofen, S.; Dörpinghaus, J.; adn Thomas Reiser, M.T.; Steiner, P. Reinventing Historical Sources as New Computational Social Science Data: Regulations for Vocational Education over Time in Germany. In Proceedings of the Digital Humanities 2024: Book of Abstracts, Graz, Austria, 14 July 2024.
39. Trufanov-Nok/Scantailor-Universal: ScanTailor Universal—A Fork Based on Enhanced+Featured+Master Versions of ST. Available online: <https://github.com/trufanov-nok/scantailor-universal> (accessed on 7 October 2024)
40. Unpaper/Unpaper: A Post-Processing Tool for Scanned Sheets of Paper. Available online: <https://github.com/unpaper/unpaper> (accessed on 7 October 2024)
41. Dörpinghaus, J.; Jacobs, M. Semantic Knowledge Graph Embeddings for biomedical Research: Data Integration using Linked Open Data. In *Proceedings of the SEMANTiCS (Posters & Demos)*; Fraunhofer: Hamburg, Germany, 2019.
42. Dörpinghaus, J.; Jacobs, M. Knowledge detection and discovery using semantic graph embeddings on large knowledge graphs generated on text mining results. In Proceedings of the 2020 15th Conference on Computer Science and Information Systems (FedCSIS), Sofia, Bulgaria, 6–9 September 2020; pp. 169–178.
43. Dörpinghaus, J.; Samray, D.; Helmrich, R. Challenges of automated identification of access to education and training in Germany. *Information* **2023**, *14*, 524.
44. Derksen, F.; Dörpinghaus, J. Digitalization and Sustainability in German Continuing Education. In *INFORMATIK 2023—Designing Futures: Zukünfte Gestalten*; Gesellschaft für Informatik e.V.: Bonn, Germany, 2023; pp. 1945–1953. https://doi.org/10.18420/inf2023_194.
45. Fischer, A.; Dörpinghaus, J. Web Mining of Online Resources for German Labor Market Research and Education: Finding the Ground Truth? *Knowledge* **2024**, *4*, 51–67.
46. Dörpinghaus, J.; Schaaf, S.; Fluck, J.; Jacobs, M. Document clustering using a graph covering with pseudostable sets. In Proceedings of the Computer Science and Information Systems (FedCSIS), Prague, Czech Republic, 3–6 September 2017; pp. 329–338.
47. Dörpinghaus, J.; Schaaf, S.; Jacobs, M. Soft document clustering using a novel graph covering approach. *BioData Min.* **2018**, *11*, 11.
48. Dörpinghaus, J. Die Soziale Netzwerkanalyse: Neue Perspektiven für die Auslegung biblischer Texte? *Biblisch Erneuerte Theol.* **2021**, *5*, 75–96.
49. Fechner, R.; Dörpinghaus, J.; Firll, A. Classifying industrial sectors from German textual data with a domain adapted transformer. In Proceedings of the 2023 18th Conference on Computer Science and Intelligence Systems (FedCSIS), Warsaw, Poland, 17–20 September 2023; pp. 463–470.
50. Fechner, R.; Dörpinghaus, J. No Train, No Pain? Assessing the Ability of LLMs for Text Classification with no Finetuning. In Proceedings of the Position Papers of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS), Belgrade, Serbia, 8–11 September 2024; pp. 9–16.
51. Dörpinghaus, J.; Klein, J.; Darms, J.; Madan, S.; Jacobs, M. SCAIView-A Semantic Search Engine for Biomedical Research Utilizing a Microservice Architecture. In *Proceedings of the SEMANTiCS (Posters & Demos)*; Fraunhofer: Hamburg, Germany, 2018.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.