information

Article

Challenges of Automated Identification of Access to Education and Training in Germany

Jens Dörpinghaus [1,2]*, David Samray [1] and Robert Helmrich [1]

[1] Federal Institute for Vocational Education and Training (BIBB), 53113 Bonn, Germany
[2] Department 4: Computer Science, University of Koblenz, 56070 Koblenz, Germany
* Correspondence: jens.doerpinghaus@bibb.de

Abstract: The German labor market relies heavily on vocational training, retraining, and continuing education. In order to match training seekers with training offers and to make the available data interoperable, we present a novel approach to automatically detect access to education and training in German training offers and advertisements and identify open research questions and areas for further research. In particular, we focus on (a) general education and school leaving certificates, (b) work experience, (c) previous apprenticeship, and (d) a list of skills provided by the German Federal Employment Agency. This novel approach combines several methods: First, we provide technical terms and classes of the education system that are used synonymously, combining different qualifications and adding obsolete terms. Second, we provide rule-based matching to identify the need for work experience or education. However, not all qualification requirements can be matched due to incompatible data schemas or non-standardized requirements such as initial tests or interviews. Although there are several shortcomings, the presented approach shows promising results for two data sets: training and retraining advertisements.

Keywords: educational data mining; education evaluation; rule-based system; computational sociology; labour market research

check for updates

Citation: Dörpinghaus, J.; Samray, D.; Helmrich, R. Challenges of Automated Identification of Access to Education and Training in Germany. *Information* **2023**, *14*, 524. https://doi.org/10.3390/info14100524

Academic Editor: Theodora Tsikrika

Received: 4 August 2023
Revised: 25 August 2023
Accepted: 14 September 2023
Published: 26 September 2023

1. Introduction

The German labor market is dynamic: technical innovations and social changes lead to new skills needed by employees. However, the German education system has a large share of vocational education and training (VET), retraining and continuing vocational education and training (CVET), which are key to responding to these new requirements [1,2]. In the Federal Republic of Germany, the Vocational Training Act (BBiG) of 1969, which was reformed in 2005, was passed for this purpose [3]. The Federal Institute for Vocational Education and Training (BIBB), which was founded in 1970 on the basis of the BBiG, prepares the content of the training regulations. In this paper we will focus on current data.

The central importance of continuing vocational training and lifelong learning has been increasingly addressed by policymakers in the context of the National Continuing Education Strategy ('Nationalen Weiterbildungsstrategie', 2019, see e.g., [4]). Here, continuing education is described as a central prerequisite for securing a skilled workforce, for ensuring the employability of all employees and thus also for national competitiveness and innovation.

As mentioned above, the German education system offers different pathways for professional qualification, see [5] and Figure 1, and we need to distinguish between initial (training, "Ausbildung" or retraining, "Umschulung") and continuing vocational education ("berufliche Weiterbildung"), which includes unregulated continuing vocational education and regulated upgrading training, which is leading to higher professional degrees ("Aufstiegsfortbildung" (BBiG/HwO, by federal states or in the health sector, see [6,7]). Here we find four different stakeholders: (a) educational institutions, (b) companies and

enterprises, (c) employees and (d) financiers. They all have to react to emerging situations, e.g., the advancing digitalization [2] and the development towards a sustainable economic system [8], by a corresponding further development of the offered vocational education contents. From a research perspective, the task is to determine which vocational education content is increasingly being offered and demanded in order to draw conclusions about the development needs of the vocational education system. The research-based further development of the vocational education system should not only ensure the competitiveness of the economy on a systemic level, but also contribute to counteracting unemployment and stabilizing the social security system [9].

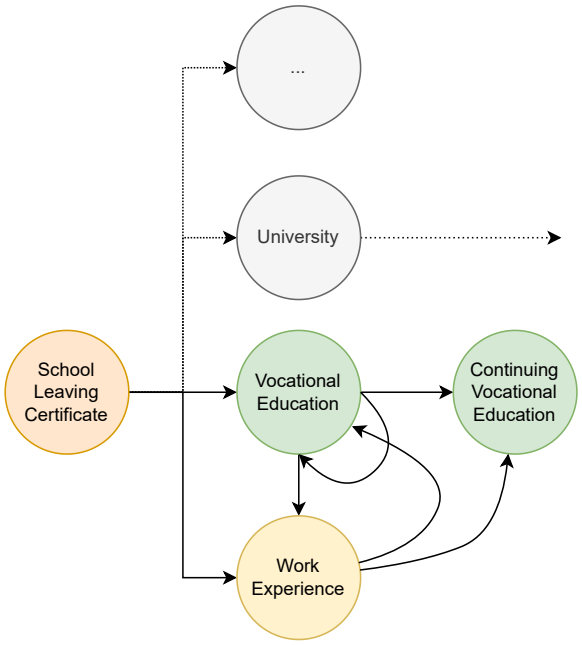

Figure 1. A simplified version of the analysis of vocational education and training in Germany. We exclude all other educational pathways, such as university programs. After leaving school, a person can enter a vocational training program and, together with work experience, enter continuing education programs. We focus here on continuing education programs because they offer the broadest formulation of access restrictions: Work experience, prior education, and educational attainment. However, this illustration covers only a kind of ideal type. For example, it is not impossible to enter a continual vocational training program without a initial vocational training, and work experience restrictions may be waived.

As a first step, we consider the automated detection of access to education and training. This will help to match education seekers and offers. However, the offers and advertisements are usually provided as free text, and we can identify two challenges: (a) these texts may be incomplete or imprecise, (b) there is no unique taxonomy or list of occupations, training, industries, and skills.

In this paper, we present a rule-based approach to detect prerequisites for access to education and training in German training advertisements and offers. In particular, we focus on (a) general education and school leaving certificates, (b) work experience ("Berufserfahrung"), (c) a previous apprenticeship, and (d) a list of skills provided by the German Federal Employment Agency (Bundesagentur für Arbeit—BA), see Figure 1. Although we can identify several shortcomings, it offers promising results for two data sets (training and retraining).

This paper is divided into six sections. The first section provides an introduction, the second section gives a brief overview of the state of the art and related work. The third section describes the methodological background, the data, the pipeline, and the matching

approaches. The fourth section is dedicated to experimental results and evaluation of this novel approach. Our conclusions and outlook are given in the final section.

2. Related Work

We find an increasing interest in mining data from educational databases, advertisements and information systems, see [10–12]. Here, supporting decision-making and the process management within education is key. The generic challenges are usually the automated extraction of knowledge from data, usually interpreted passages from texts, and the mapping to existing data sets. However, there are still several challenges on data and data integration, see [13]. We will now focus on the different data sets: general school and education degrees and school-leaving certificates, professional experience, a previous apprenticeship and skills.

Skill concepts have been heavily used for the analysis of job advertisements (job ads), see [14], and their visualization, see [15]. They offer a good starting point for matching open positions to corresponding employees. The proposed technologies range from automated mapping of search terms to classification of skills, see [16]. In education research and pedagogy, several approaches towards competencies and skills exist which provide divergent resources: Not only with different meaning but also with different synonyms. While in English Competences, Skills and Knowledge are often used as synonyms, this does not hold for German language: Some words refer to multiple concepts, while some concepts are labeled with multiple words, see [17].

Several approaches have focused on extracting professional experience from texts. However, they are limited to English texts, e.g., resumes, see for example [18,19]. The first challenge is that neither in English nor in German the concept of professional experience is clearly defined. Next, while according to our knowledge only approaches for English texts exist, a rule-based approach seems to work best, especially for initial research.

Similarly, previous apprenticeship are usually considered only in the context of job ads [20,21] and for general school and education degrees and school-leaving certificates no approaches on German texts exist to the best of our knowledge.

There is also very little research beyond the classical methods. We find some recent publications on transformer models [22–24]. In general, all methods have their own biases and assumptions that need to be carefully considered, see for example [25–27].

In summary, not only does this field encompass a wide variety of related issues, but it has not yet received the scholarly attention it deserves. This may be due to the highly interdisciplinary nature of the field: While scholars in education or the social sciences focus on other aspects, computer scientists are interested in other problems. In addition, labor market research is highly regional, and most resources are limited to a particular language.

Since we can only rely on very limited previous work, we will continue with a detailed discussion of the methods to encourage further research in this field.

3. Method

3.1. Labor Market Data: Occupations and Skills

Labor markets are complex fields with diverse data structures and multiple applications, for example, connecting jobseekers to the right training or job [28]. For this reason the multilingual classification of European Skills, Competences, Qualifications and Occupations (ESCO) is a good example for the central role of ontologies in this field, see [29]. However, ESCO cannot provide all details of local labor market needs and does not provide links to other hierarchies of skills. For example, in German-speaking countries, other taxonomies of occupations and skills are widely used. Thus, when discussing data for occupational qualifications and certificates, we need to consider multiple data sets.

The International Standard Classification of Occupations (ISCO) (See https://www.ilo.org/public/english/bureau/stat/isco/isco08/, checked 20 March 2023) 2008 was developed by the International Labour Organization (ILO) and was published in 1958, 1968, 1988 and as its recent version in 2008. It was also used within the European Union (EU),

and some German-speaking countries (Germany, Austria, Switzerland) have developed a specific version of the ISCO 2008. ISCO also maps to the ontology "European Skills, Competences, Qualifications and Occupations" (ESCO) which also adds skills and competences to ISCO.

In Germany, the Classification of Occupations ("Klassifikation der Berufe", KldB) (See https://statistik.arbeitsagentur.de/DE/Navigation/Grundlagen/Klassifikationen/Klassifikation-der-Berufe/KldB2010-Fassung2020/KldB2010-Fassung2020-Nav.html, accessed on 1 August 2023) as part of the DKZ (Documentation number, "Dokumentationskennziffer") is the reference for IAB (Institut für Arbeitsmarkt- und Berufsforschung) and the German Federal Employment Agency (Bundesagentur für Arbeit-BA). The most recent version is the 2020 revision of KldB 2010, which was completely redeveloped and makes previous versions from 1988 and 1992 deprecated. It was developed to be compatible to ISCO-08. This data are part of the matching process in the BA and are integrated into other IT applications. However, while part "B" of KldB is dedicated to occupations, part "C" covers continuing vocational education, "K" skills, and "A" higher education. All these parts are important to describe the access to education and training, see Table 1 for details.

Table 1. Data sets used to describe the access to education and training.

Data Set	Content	Entries
A	Higher Education	797
B	Occupations	33,802
C	Continuing Vocational Education	542
K	Skills	9078

Part A mainly describes general academic education. These data are complex because each federal state in Germany has a different approach, e.g., vocational high schools ("Berufliche Gymnasien") do not exist in every federal state. In addition, some older terms are often used, e.g., secondary school diploma ("Realschulabschluss") instead of secondary degree I ("Sekundarabschluss I"). More precisely, 23 terms in A use "Realschulabschluss" as a synonym. Mapping these 797 entries to the textual description will be a first challenge. The complexity of the German school leaving certificate is a topic in itself, see [30]. However, since we only focus on the correct mapping of natural language and DKZ-A, we will map each special certificate to its identifier, e.g., evening high school, vocational high school/technical high school, technical college ("Abendgymnasium, Berufsoberschule/Fachoberschule, Fachschule") to the same, and the broader terms, e.g., advanced technical college entrance qualification ("Fachhochschulreife") to all references, see Section 3.4.

Part B provides a comprehensive description of 33,802 occupations ordered by occupational areas (e.g., 1 Occupations in agriculture, forestry, farming, and gardening), occupational main group (e.g., 12 Occupations in gardening and floristry), occupational group (e.g., 122 Occupations in floristry), occupations sub-group (e.g., 1229 Supervisors and managers in floristry) and occupational types (e.g., 12,294 Managers in floristry). However, not all of these are relevant as we focus on initial and continuing vocational education and training. Therefore, we will add generic terms to describe the need for prior education (B+) and work experience (BE). For the same reason, part C is added for the sake of completeness. For example, very few training courses will only focus on master craftsman or master craftswoman.

We will now describe some existing taxonomies for skills in the German language and discuss why we limit our approach to BA data. Our first example is the European Classification for Skills, Competences, Qualifications and Occupations (ESCO). It provides a multi-language hierarchy of skills and competences (and in addition qualifications and occupations) containing a full text description, scope notes and comprising examples. Gonzalez et al. state, that only few works have described the analysis and use of ESCO,

see [31]. Some work has been carried out for semantic interoperability between skills and labour market documents, which was initially promised by ESCO [32]. Other scholars tried to use data from ESCO and Wikidata for text mining on scientific literature, see [31] or for curriculum analytics, see [33]. Recent research has provided a generic mining and mapping approach [34] and automated ontology alignment for ESCO and the English O*NET [35].

But ESCO is not the only skill systematic available in German language, as we will discuss now. Krebs et al. provide a brief overview about the challenges when working with these resources [17]. First, the competencies must be relevant to work. Competencies and skills must be at a level of abstraction which can be used to distinguish between jobs. This means that competencies needs to be defined so broadly that they usually appear in more than one individual workplace. Competencies must offer added value compared to other existing classifications, which means they must be multidisciplinary—but at the same time, however, they should not be limited to a pure measurement of cognitive skills, as already implied in the qualification or requirement level. Obviously, again, this highlights the interdisciplinary challenges: A mapping to skills is not only a technical, semantic link between two resources, but should also keep the information and abstraction layers.

While ESCO is funded by the EU and openly available, other resources are more restricted in their usage. For example, both employment agencies in Germany and Austria provide their own resources for structuring skills and linking them to occupations: the BERUFENET is developed by the German Federal Employment Agency (BA) and the 'Klassifikation des Arbeitsmarktservice' is provided by the Austrian Labour Market Service (AMS). Both taxonomies are available via an online platform and are not interoperable, see [17].

Another concept of German skills was introduced by MYSKILLS, see [36,37]. It is primarily intended to define and discuss skills needed for a particular profession in the context of "Unlocking the Potential of Migrants in Germany" [38]. It contains a hierarchical definition of skills for ca. 30 professions as unstructured resources (data is available as PDF). The skills have an identifier and a description. Skills are also used in the field of adult training and lifelong learning. The model GRETA is special, because it also introduces several levels of skills, see [39,40].

As we can see, all different concepts have (a) different application purposes, (b) different data structures, (c) a different level of machine-readability and (d) are in general not interoperable. Several attempts were made for a standardization in this field. Konert et al. provided a different perspective for the modelling of competences, see [41]. They underlined validity and reliability as primary challenges for skill definitions. Rentzsch and Staneva provided a play for combining taxonomies and ontologies with data-driven methods in the field of skills matching and skills intelligence, see [42]. Focusing on the modelling of skills, Ref. [43] introduced HyperCMS, an approach for semantic modelling in knowledge graphs.

Summarizing, for this version we decided to use the skills from BERUFENET, because they are included in the data used for BERUFENET database (see below) and form the most generic list. However, for future versions, including generic skills from ESCO and AMS will be key.

3.2. Data

Our first evaluation data was obtained from the data portal KURSNET of BA (See https://www.arbeitsagentur.de/kursnet, accessed 10 September 2023). We retrieved data for initial and continuing vocational education and training, which leads to 23,460 (continuing professional education), and 66.549 (re-training) results in February 2023. Within the advertisement text, we only considered the section access information ("Zugangsinformationen") with the subsection access ("Zugang") which usually holds natural language describing the access limitations, see Figure 2.

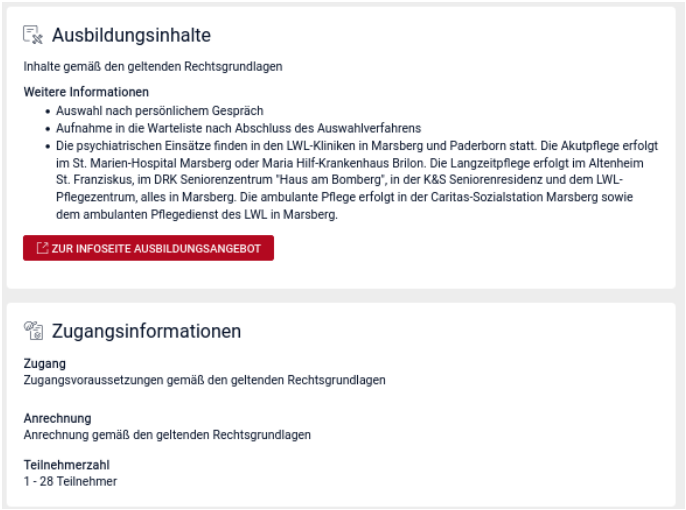

Figure 2. A screenshot of KURSNET data entries with two sections: training content ("Ausbildungsinhalt") and access information ("Zugangsinformationen") with the subsection access ("Zugang").

All data sets have a particular usage of terms from the previously described sets, see Table 2. These texts may refer to specific and unspecific access restrictiveness. For example:

> Personal counseling interview, PROFILE aptitude test, at least secondary school leaving certificate required
> (Persönliches Beratungsgespräch, PROFIL-Eignungstest, mind. Hauptschulabschluss erforderlich)

Table 2. Overview of data to evaluation the approach and their characteristic data content. ED refers to general school and education degrees and school-leaving certificates, PE to professional experience, PA to previous apprenticeship and S to skills. Here, + indicates a wide usage of entities, - to an average usage and — to nearly no usage.

Data Set	Source	Size	ED	PE	PA	S
Re-trainings	KURSNET	120	+	+	-	-
Continuing professional education	KURSNET	120	+	+	—	-
Continuing professional development	Weiterbildungsportal RLP	120	-	+	+	+

This text contains two unspecific access restrictions, a counselling interview, a non-standardized aptitude test and a lower secondary school-leaving certificate ("Hauptschulabschluss"). Other texts do not describe a specific access restriction at all:

> good general education, manual dexterity, ability to paint, nationwide mobility
> (gute Allgemeinbildung, handwerkliches Geschick, Farbtauglichkeit, bundesweite Mobilität)

No standardized qualifications for general education, being good with hands, color blindness or mobility exist. This text cannot be mapped to the data described above. However, we selected 120 advertisements each to create a gold standard. This curation process was done manually.

A second dataset to evaluate the proposed approach was obtained from "Weiterbildungsportal Rheinland-Pfalz" (See https://weiterbildungsportal.rlp.de/, accessed 10 September 2023). Here, the data mainly focuses on continuing professional development. See Table 2 for a detailed overview on how this data set differs from the previously discussed. Thus, the access information is slightly different and often require one particular vocational education:

> Successfully completed training in a recognized three-year commercial apprenticeship in the trade ...

(Eine mit Erfolg abgeschlossene Ausbildung in einem anerkannten dreijährigen kaufmännischen Ausbildungsberuf im Handel ...)

While this data is not in the primary focus of this work, which mainly focuses on trainings and re-trainings, it will show if our approach generalizes.

For testing purposes, we manually annotated a set of 120 training (Berufsausbildung) and re-training (Umschulung) advertisements text snippets for each dataset. While higher education, school-leaving certificates, and prior work experience are usually not really interpretable, skills can be operationalized in different ways, and thus this evaluation set has its own biases. We will discuss this in more detail in Section 4. In general, for this very complex topic, we refer to further discussion in [22]. Since the data is copyrighted, we can only publish a small part of the evaluation.

3.3. Workflow

The workflow comprises two independent parts, building the mappers and processing the corpora data, see Figure 3. All parts have been developed using Python 3.9.2, spaCy 3.3.0, an advanced NLP library, and beautifulsoup4 4.9.3. Extracting the section from KURSNET website was done by extracting the content of the <p>-tag. However, we publish a generic method that can be used on any textual data. The pre-processing is done using spaCy using a tokenizer, sentence splitter, NER and the matcher. We use three different matchers to detect access information in the text. We use a *PhraseMatcher* to match the large lists of entities described in sets B, C and K. For A, we build additional mappings and add synonyms. However, for additional rule-based matching, we use the *Matcher* class. For more technical details, we refer to [44] or the spaCy documentation.

We will now continue with a detailed discussion of how to map synonyms and qualifications in data sets A and B and how to build the additional rules.

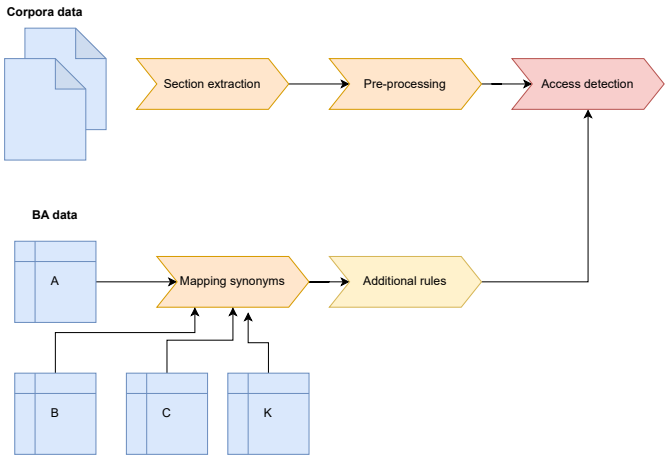

Figure 3. Flow-diagram of the processing workflow. The letters correspond to DKZ parts introduced in Section 3.1.

3.4. Mapping Synonyms in Education

Usually, together with KldB, only five different school leaving certificates are considered relevant for labor market research [45]: Without school-leaving certificate, Secondary/elementary school leaving certificate, Secondary school leaving certificate or equivalent, High school diploma/general university entrance qualification, Degree unknown. However, as discussed above, they have different names in different states and we need to consider the appropriate mapping.

Therefore, we will now describe the algorithmic approach to mapping synonyms in data sets A and B containing a list of general academic education and occupations, see Algorithm 1. The data in A, obtained from the BA, usually describes a general school and an educational degree, in this example “Vocational High School, Technology (General

University Entrance Qualification)" and "Vocational High School, Economics (General University Entrance Qualification)":

A 1.SAN-21.1 | Berufliches Gymnasium, Technik (allgemeine Hochschulreife) |
A 1.SAN-21.2 | Berufliches Gymnasium, Wirtschaft (allgemeine Hochschulreife) |

The first part of each line contains the ID of a term (e.g., "A 1.SAN-21.1"), the second part the information. So the first rule splits the degree (in this case "Allgemeine Hochschulreife") and the school. See the line 8 in the Algorithm 1. However, some terms list only courses that refer to a preparation for education ("Vorbereitung"), which must be omitted since they do not lead to a formal qualification, see for example "Preparation for educational courses leading to the general higher education entrance qualification" and "Preparation for the examination for the general higher education entrance qualification":

A 1.SAN-91 | Vorbereitung auf Bildungsgänge, die zur allgemeinen Hochschulreife führen |
A 1.SAN-92 | Vorbereitung auf die Prüfung zur allgemeinen Hochschulreife |

In addition, we provide an extensive list of synonyms as described above, e.g., "mittlerer Schulabschluss" to "Realschulabschluss", "Fachoberschulreife", "Sekundarabschluss I" and "mittlere Reife". Some terms are obsolete but still in common use. In the line 3 in the Algorithm 1 additional synonyms are added.

The situation is more complex when focusing on record B. Here, male and female titles are provided together, and optional regular information is provided. For example:

B 27302-902 | Produktionstechnologe/Produktionstechnologin |
B 28222-905 | Fachpraktiker/Fachpraktikerin für Näherei und Schneiderei (§66 BBiG/§42r HwO) |

In the first case, a complete male and female title is split by a slash. In the second case, only the first part of the title is split. In line 12 the two functions *male* and *female* are used to describe the extraction process.

Algorithm 1 MAPPING SYNONYMS

Require: List L of entries from data sets A and B
Require: List S of synonyms for general school and education degrees
Ensure: List L' with synonyms
$L' = \emptyset$
for every $l \in L$ **do**
  for $s \in S$ **do**
    if $s \in l$ **then**
      $L' \leftarrow [id(l), s]$
    end if
  end for
  if "(" $\in l$ **then**
    $L' \leftarrow [id(l), split(l, "(", 0)]$
    $L' \leftarrow [id(l), split(l, "(", 1).remove(")")]$
  end if
  if "/" $\in l$ **then**
    $L' \leftarrow [id(l), male(l)]$
    $L' \leftarrow [id(l), female(l)]$
  end if
end for
return $L \cup L'$

3.5. Rule-Based Matching

The complexity of natural text leads to several challenges. We find multiple ways to describe professional experience ("Berufserfahrung"), school leaving certificates ("Schulabschluss") or technical and vocational education and training.

Professional experience is mentioned in several more or less complex sentences:

But also dropouts or persons **with work experience .** in other fields should inform themselves about the retraining offer if they are interested...
Completed vocational training or **work experience**...
work experience desirable (1 year), but not a must....
(Aber auch Studienabbrecher/innen oder Personen **mit Berufserfahrung** in anderen Bereichen sollten sich bei Interesse über das Umschulungsangebot informieren...
Abgeschlossene Berufsausbildung oder **Berufserfahrung**...
Berufserfahrung wünschenswert (1 Jahr), aber kein Muss...)

In the last sentence, experience is only preferable ("wünschenswert"), but not a condition. School-leaving certificates are usually found in a similar context and setting. However, as described above, since particular certificates are usually named explicitly the extraction itself is not that challenging. See Figure 4 (bottom) for an illustration of a rule matching several complex phrases. However, extracting previous vocational education is more challenging. Some texts directly mention the prerequisite:

Minimum requirement: **vocational training**
(Mindestvoraussetzung: **Berufsausbildung**)

Usually, a vocational education is embedded in an or-clause stating that it is not the only way to access the training:

If possible, completed **vocational training** or ...
At least 4 years of professional activity or completed **vocational training**....
(Möglichst abgeschlossene **Berufsausbildung** oder ...
mind. 4 Jahre berufliche Tätigkeit oder eine abgeschlossene **Berufsausbildung**...)

However, a sentence matched by a rule for negative phrases (... **keine** abgeschlossene Berufsausbildung...) might also be matched by a positive rule (... eine abgeschlossene Berufsausbildung...). Thus, we define positive and negative results (B+, B-) and an existing negative result will replace a positive one, see Figure 4 (top) for an illustration.

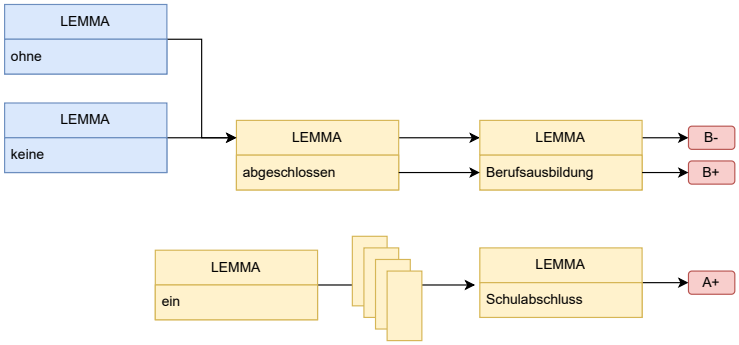

Figure 4. Two example rules to detect school-leaving certificates (Schulabschluss, **bottom**) or technical and vocational education (**top**). All rules have positive (e.g., A+) and negative (e.g., A-) results. An existing negative result will remove a positive one.

We provide several rules for frequently used structures like "with our without" (... ohne oder mit...) or negative clauses. Our rule-based approach does not differentiate between or- and and-clauses. However, we observed that most descriptions use or-clauses. Thus, in these cases, all detected qualifications give access to the training offered. Another limitation is the naming of language qualifications ("angemessene Deutschkenntnisse" or "Deutschkenntnisse ... B2"). The biggest challenge is that these qualifications cannot be exactly matched in KldB which only offers "Deutsch-verschiedene Niveaustufen" (A 8.11) or Grundstufe, Mittelstufe and Oberstufe. Second, we only found requirements for German, not for any other language. Thus, as a preliminary solution, we map these rules to "A 8.11".

3.6. Limitations

As described above, we rely heavily on the BA data. While school qualifications and occupations are regulated by the German government, skills and other tests are usually not clearly defined, and the BA data omit several of them. We decided to leave the missing data for this version, as they require more analysis of their nature. For example, we found "ZAUG-Eignungsfeststellung" or "CDT-Eignungstest für IT-Berufe". Thus, unspecific access restrictions, e.g., a counseling interview, and non-standardized tests are currently not considered.

4. Experimental Results

As described in Section 3.2 we rely on data from BERUFENET database for initial and continuing vocational education and training. Our gold standard to evaluate the output of our approach are 120 sets for training (Berufsausbildung) and re-training (Umschulung) advertisements. A third data set focusing on continuing professional development was obtained from "Weiterbildungsportal Rheinland-Pfalz". To analyze the quality of this network, we can calculate precision and recall:

$$Precision = \frac{TP}{TP + FP}, \quad Recall = \frac{TP}{TP + FN}$$

Here, TP refers to true positive results, FP to all false positive results, and FN are false negative results. With precision and recall, we can compute the F_1-score, which is as a weighted average of the precision and the recall:

$$F_1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

For a detailed overview, see Table 3. From this table we can see that for the first two data sets the F_1 score is high. What is interesting in this data is that the precision for the retraining data is much lower than for the training data. This is mainly due to misrecognition of occupations. For example, one advertisement asks people to call the consultant if they have any questions:

> ... and have their personal access requirements checked by our **consultants**.
> (... und ihre persönlichen Zugangsvoraussetzungen durch unsere **Fachberater/innen** prüfen lassen.)

Table 3. Overview of metrics and results: Precision, recall and F_1 scores on different data sets.

Data Set	Precision	Recall	F_1-Score
Re-trainings	0.86	0.95	0.91
Training	0.99	0.95	0.97
Continuing professional development	0.47	0.30	0.36

However, our approach recognizes several professions to access this training: "Fachberater/Fachberaterin für Softwaretechniken/Finanzdienstleistungen/Vertrieb/integrierte Systeme" and even another position in commerce: "Fachverkäufer/Fachberater/Fachverkäuferin-Bau-/Heimwerkerbedarf". Another difficulty is the legal aspect. For example, access to training may be easier with support (so-called "Bildungsgutschein"), so this is mentioned in the advertisement: "SGB II/SGB III (Bildungsgutschein)". In this case, "SGB III" also refers to the unemployment insurance and the skill 030400-011 (Sozialgesetzbuch III) is recognized. These two examples directly illustrate the limitations of string-matching approaches, which could be addressed with transformer models, for example. However, the recall is generally high.

Interestingly, the results change for data set three. When comparing the results with the other data sets, neither precision nor recall show a working approach. In this data, prior training is needed more often. We will start with a complex example:

> Successful **completion of training** in a **recognized three-year commercial trade occupation** or at least one year's professional experience, or successful **completion of the final examination for sales assistant** or in another **recognized trade occupation**, followed by at least five years' professional experience if the above requirements are not met.
> (Eine mit Erfolg **abgeschlossene Ausbildung** in einem anerkannten dreijährigen **kaufmännischen Ausbildungsberuf im Handel** oder eine mindestens einjährige Berufspraxis oder eine mit Erfolg abgelegte **Abschlussprüfung zum Verkäufer/zur Verkäuferin** oder in einem anderen **anerkannten Ausbildungsberuf** und danach eine mindestens fünfjährige Berufspraxis, wenn die zuvor genannten Voraussetzungen fehlen.)

This text describes the following requirements: An apprenticeship in a trade, *or* one year of work experience, *or* an apprenticeship as a shop assistant, *or* any other apprenticeship *and* five years of work experience. However, while there are few shop assistant apprenticeships, it is unclear which apprenticeships are in trade (either "business" or "commerce"–"kaufmännisch" or "im Handel"). In any case, our approach fails to enumerate them precisely. Thus, further research also needs to cover the hierarchy of occupations, although there will always be room for ambiguity. Logical clauses also need further attention. In general, we did not find any reference to a specific training, but only general statements:

> A **completed electrotechnical training** oriented to the specifications of TRBS 1203.... A completed **technical vocational training**...
> (Eine **abgeschlossene elektrotechnische Ausbildung** orientierend an den Vorgaben der TRBS 1203...
> Abgeschlossene **technische Beraufsausbildung**...)

This shows that our approach is able to detect some elements to access education and training in German language: The mapping of synonyms in education works well and extracts general school and training qualifications required for a specific training. The rule-based matching approach to detect work experience, previous apprenticeship and school qualifications and to extract skills from the BA list shows promising results. However, the extraction of specific occupations and education does not work as expected. Our approach either extracts occupations that are not a prerequisite or cannot dissolve groups of trainings.

Thus, for further research, more contextual data (hierarchies or taxonomies, e.g., of occupations) should be considered, together with extraction methods that can exploit the context of entities, e.g., for certain legal aspects or other lemmata that remain ambiguous. We will discuss more of these aspects for further research in the next section. However, it is striking that the performance is highly dependent on the ads considered. Thus, a careful assessment of what is actually part of the extracted data and what is contained in the textual resources is key for further research.

In summary, the proposed method works well for data that does not use occupational or training references, namely training and pre-training advertisements, but does not yet generalize.

5. Conclusions and Outlook

Vocational education and training, retraining, and continuing education are key to meeting the demands of tomorrow's labor market. In order to match training seekers with training offers and to generate interoperable data, we presented a novel approach to automatically detect access to education and training in German training offers and advertisements.

We focused on (a) general education and school leaving certificates, (b) work experience, (c) previous apprenticeship and (d) skills. Our novel approach combined several methods: First, we provided a mapping of synonyms in education, combining different qualifications and adding obsolete terms. As our results and discussion show, this works well for the evaluation data.

Second, we provided rule-based matching to identify the need for professional experience or apprenticeship. As discussed, not all entry requirements can be matched due to incompatible data schemas or non-standardized requirements, such as initial tests or interviews. Extracting specific occupations and education does not work as expected. Our approach either extracts occupations that are not prerequisites or cannot resolve groups of trainings.

While our approach only applied rule-based and string matching methods, it shows that further research is needed in this area. The presented approach shows that, on the one hand, the initial data to be mapped and, on the other hand, the text structures have a great impact on the efficiency and accuracy of the presented approach. The experimental results show promising results for two data sets (training and re-training), but further research is needed, especially on the identified shortcomings, in order to build a generic tool suitable for different input data. For data, we also need to consider other types of data. DKZ of the BA offers other data, which are not used yet. For example, H "Higher Education Programs" or P "Occupational Medicine and Occupational Psychology Characteristics", which would help to improve the results. The addition of other data sources such as ESCO poses further challenges: In most cases, concepts such as skills or job tasks are fuzzy, and it remains unclear how they have been operationalized, which means that data linkage often requires researchers to manually compare data and other contextual factors in order to combine data sets. We also discussed that skills in particular are difficult to operationalize. Therefore, further research should also evaluate possible biases in ad text. For this purpose, we have provided a subset of our evaluation dataset for the scientific community.

Further research could also investigate how other methods, such as fine-tuned pre-trained transformer models, would help to extract the relevant data. However, it is still necessary to define a set of interoperable data as 'ground truth', otherwise the matching process could not be used to match workers and opportunities. Thus, the data perspective currently presents the greater hurdles for further research.

However, while matching training seekers and offers is the most natural application, we have integrated the presented approach into a generic workflow for analyzing labor market data in a knowledge graph, see Figure 5. We propose that further data integration (e.g., of other skill taxonomies) is needed to model and analyze training ads. This approach to understanding access to education and training in Germany is only a first step.

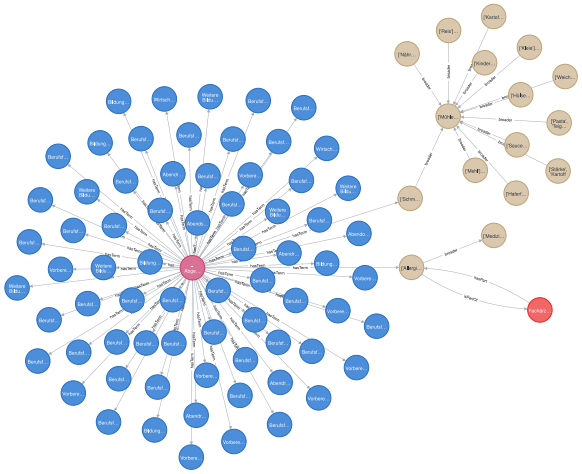

Figure 5. *Cont.*

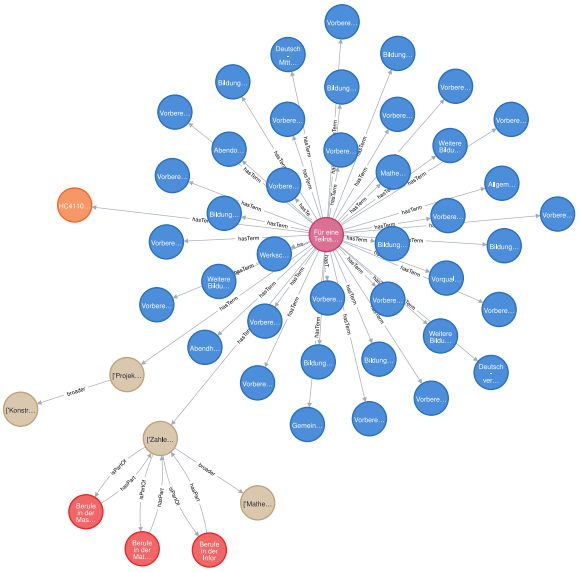

Figure 5. Two example outputs of the proposed workflow embedded in the knowledge graph of labor market data. The training ad (purple) is in the center, blue nodes are related to education, brown nodes to skills, orange to vocational training. The skills can be linked to occupations (red).

Author Contributions: Conceptualization, J.D., R.H. and D.S.; methodology, J.D. and D.S.; software, J.D.; validation, J.D. and D.S.; formal analysis, J.D. and D.S.; investigation, J.D. and D.S.; resources, J.D. and D.S.; data curation, J.D. and D.S.; writing—original draft preparation, J.D.; writing—review and editing, J.D. and D.S.; visualization, J.D.; supervision, J.D. and R.H.; project administration, J.D.; funding acquisition, R.H. 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: Since most of the input data is copyrighted, we can only publish a small set of the evaluation data, but all models which are available at https://github.com/TM4VETR/DetectAccessToEducation (accessed on 10 September 2023).

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

Abbreviations

The following abbreviations are used in this manuscript:

VET	Vocational Education and Training
CVET	Continuing Vocational Education and Training
DKZ	Documentation number, "Dokumentationskennziffer"
BA	German Federal Employment Agency (Bundesagentur für Arbeit)
B+	prior education
BE	work experience
KldB	German Classification of Occupations, "Klassifikation der Berufe"

References

1. Dobischat, R.; Käpplinger, B.; Molzberger, G.; Münk, D. Digitalisierung und die Folgen: Hype oder Revolution? In *Bildung 2.1 für Arbeit 4.0*? Springer: Wiesbaden, Wiesbaden, 2019; pp. 9–24.
2. 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, BIBB: Bonn, Germany, 2016.
3. Kuppe, A.M.; Lorig, B.; Schwarz, H.; Stöhr, A. *Ausbildungsordnungen und wie sie Entstehen*; Bundesinstitut für Berufsbildung: Bonn, Germany, 2015.
4. Schiersmann, C. Weiterbildungsberatung im Kontext der Nationalen Weiterbildungsstrategie: Finanzielle und strukturelle Aspekte. *Hessische Bl. Volksbild.* **2022**, *72*, 43–53. [CrossRef]

5. 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*; Gonon, P., Bürgi, R., Eds.; Peter Lang: Bern, Switzerland, 2021.
6. Dikau, J. Rechtliche und organisatorische Bedingungen der beruflichen Weiterbildung. *Handb. der Berufsbild.* **1995**, 427–440.
7. Bauer, R.; Bauer, R. Die Debatte über die Zukunft der dualen Berufsausbildung. In *Verberuflichung von Weiterbildung und die Zukunft der Dualen Berufsausbildung: Eine Berufssoziologische Analyse am Beispiel des Kraftfahrzeuggewerbes*; Springer: Wiesbaden, Germany, 2000; pp. 21–84.
8. Steeg, S. *Die Wasserstoffwirtschaft in Deutschland: Folgen für Arbeitsmarkt und Bildungssystem; eine Erste Bestandsaufnahme*; Bundesinstitut für Berufsbildung: Bonn, Germany, 2022.
9. Zimmermann, K.F.; Biavaschi, C.; Eichhorst, W.; Giulietti, C.; Kendzia, M.J.; Muravyev, A.; Pieters, J.; Rodríguez-Planas, N.; Schmidl, R. Youth unemployment and vocational training. *Found. Trends® Microecon.* **2013**, *9*, 1–157. [CrossRef]
10. Romero, C.; Ventura, S. Educational data mining: A survey from 1995 to 2005. *Expert Syst. Appl.* **2007**, *33*, 135–146. [CrossRef]
11. Mohamad, S.K.; Tasir, Z. Educational data mining: A review. *Procedia-Soc. Behav. Sci.* **2013**, *97*, 320–324. [CrossRef]
12. Dutt, A.; Ismail, M.A.; Herawan, T. A systematic review on educational data mining. *IEEE Access* **2017**, *5*, 15991–16005. [CrossRef]
13. Kovalev, S.; Kolodenkova, A.; Muntyan, E. Educational data mining: Current problems and solutions. In Proceedings of the 2020 V International Conference on Information Technologies in Engineering Education (Inforino), Moscow, Russia, 14–17 April 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 1–5.
14. Degenhardt, S. Kompetenzen für eine digitalisierte Arbeitswelt–Anforderungen an Aus-und Weiterbildung. In *Digitaler Wandel in der Sozialwirtschaft*; Nomos Verlagsgesellschaft mbH & Co. KG: Baden-Baden, Germany, 2018; pp. 259–272.
15. Kreuzer, C. Visualisierung der Opportunity Recognition-Kompetenz von Industriekaufleuten. *Z. Berufs-Und Wirtsch.* **2018**, *114*, 247–271. [CrossRef]
16. Ziegler, P. *Zur Verwendung von Berufsinformation im Hinblick auf Matching in Deutschland und Österreich;* Technical Report; AMS info; Arbeitsmarktservice Österreich: Wien, Austria, 2012.
17. Krebs, B.; Maier, T. *Die QuBe-Kompetenzklassifikation als Verdichtende Perspektive auf Berufliche Anforderungen*; Technical Report; Wissenschaftliche Diskussionspapiere, Bundesinstitut für Berufsbildung: Bonn, Germany, 2022.
18. Ben Abdessalem, W.K.; Amdouni, S. E-recruiting support system based on text mining methods. *Int. J. Knowl. Learn.* **2011**, *7*, 220–232. [CrossRef]
19. Kopparapu, S.K. Automatic extraction of usable information from unstructured resumes to aid search. In Proceedings of the 2010 IEEE International Conference on Progress in Informatics and Computing, Shanghai, China, 10–12 December 2010; IEEE: Piscataway, NJ, USA, 2010; Volume 1, pp. 99–103.
20. Bereşewicz, M.; Pater, R. *Inferring Job Vacancies from Online Job Advertisements*; Publications Office of the European Union: Luxembourg, 2021.
21. Hermes, J.; Schandock, M. Stellenanzeigenanalyse in der Qualifikationsentwicklungsforschung. In *Die Nutzung Maschineller Lernverfahren zur Klassifikation von Textabschnitten;* Bundesinstitut für Berufsbildung: Bonn, Germany, 2016.
22. Binnewitt, J.; Krüger, K. Extracting fuzzy concepts from online job advertisements in German. In Proceedings of the 2nd Workshop on Computational Linguistics for Political Text Analysis (CPSS-2022), Potsdam, Germany, 12 September 2022; pp. 71–72.
23. Fechner, R.; Dörpinghaus, J.; Firll, A. Classifying Industrial Sectors from German Textual Data with a Domain Adapted Transformer. In Proceedings of the 18th Conference on Computer Science and Intelligence Systems, Warsaw, Poland, 17–20 September, 2023; Ganzha, M., Maciaszek, L., Paprzycki, M., Ślęzak, D., Eds.; Annals of Computer Science and Information Systems; IEEE: Piscataway, NJ, USA, 2023; Volume 35.
24. Krüger, K. Ausklasser—A classifier for German apprenticeship advertisements. In Proceedings of the Communication Papers of the 17th Conference on Computer Science and Intelligence Systems, Sofia, Bulgaria, 4–7 September 2022; Ganzha, M., Maciaszek, L., Paprzycki, M., Ślęzak, D., Eds.; Annals of Computer Science and Information Systems; IEEE: Piscataway, NJ, USA, 2023; Volume 36.
25. Nadif, M.; Role, F. Unsupervised and self-supervised deep learning approaches for biomedical text mining. *Briefings Bioinform.* **2021**, *22*, 1592–1603. [CrossRef] [PubMed]
26. Murorunkwere, B.F.; Ihirwe, J.F.; Kayijuka, I.; Nzabanita, J.; Haughton, D. Comparison of Tree-Based Machine Learning Algorithms to Predict Reporting Behavior of Electronic Billing Machines. *Information* **2023**, *14*, 140. [CrossRef]
27. Zheng, J.; Liu, Y. What does Chinese BERT learn about syntactic knowledge? *PeerJ Comput. Sci.* **2023**, *9*, e1478. [CrossRef] [PubMed]
28. Felsenstein, D.; McQuaid, R.W. Introduction to the special issue: Linking demand and supply in local labor market research. *Ann. Reg. Sci.* **2006**, *40*, 389–392. [CrossRef]
29. De Smedt, J.; le Vrang, M.; Papantoniou, A. ESCO: Towards a Semantic Web for the European Labor Market. In Proceedings of the Workshop on Linked Data on the Web co-located with the 24th International World Wide Web Conference (WWW 2015), Florence, Italy, 19 May 2015. Available online: https://ceur-ws.org/Vol-1409/paper-10.pdf (accessed on 10 September 2023).
30. Cortina, K.S.; Baumert, J.; Leschinsky, A.; Mayer, K.U.; Trommer, L. Das Bildungswesen in der Bundesrepublik Deutschland. In *Strukturen und Entwicklungen im Überblick*; Rowohlt: Reinbek, Germany, 2003.

31. González, L.; García-Barriocanal, E.; Sicilia, M.A. Entity Linking as a Population Mechanism for Skill Ontologies: Evaluating the Use of ESCO and Wikidata. In Proceedings of the Research Conference on Metadata and Semantics Research, Madrid, Spain, 2–4 December 2020; Springer: Cham, Switzerland, 2020; pp. 116–122.
32. le Vrang, M.; Papantoniou, A.; Pauwels, E.; Fannes, P.; Vandensteen, D.; De Smedt, J. Esco: Boosting job matching in europe with semantic interoperability. *Computer* **2014**, *47*, 57–64. [CrossRef]
33. Kitto, K.; Sarathy, N.; Gromov, A.; Liu, M.; Musial, K.; Buckingham Shum, S. Towards skills-based curriculum analytics: Can we automate the recognition of prior learning? In Proceedings of the Tenth International Conference on Learning Analytics & Knowledge, Frankfurt, Germany, 23–27 March 2020; pp. 171–180.
34. Fareri, S.; Melluso, N.; Chiarello, F.; Fantoni, G. SkillNER: Mining and mapping soft skills from any text. *Expert Syst. Appl.* **2021**, *184*, 115544. [CrossRef]
35. Neutel, S.; de Boer, M.H. Towards Automatic Ontology Alignment using BERT. In Proceedings of the AAAI Spring Symposium: Combining Machine Learning with Knowledge Engineering, Palo Alto, CA, USA, 22–24 March 2021
36. Köppl, C. Fachkräftemangel durch Quereinsteiger abfedern. *Wirtsch. Manag.* **2020**, *12*, 382–385. [CrossRef]
37. Rittberger, M. Digitale Bildung: Rolle und Chancen einer Forschungsinfrastruktureinrichtung. In Proceedings of the ZPID-Kolloquium 2019, Trier, Germany, 14 March 2019; ZPID (Leibniz Institute for Psychology Information): Trier, Germany, 2022.
38. Bergseng, B.; Degler, E.; Lüthi, S. Getting migrants ready for vocational education and training in Germany. In *Unlocking the Potential of Migrants in Germany*; Organisation for Economic Co-operation and Development: Paris, France, 2019.
39. Lencer, S.; Strauch, A. Ein Kompetenzmodell für Lehrende in der Erwachsenen-und Weiterbildung: Erste Ergebnisse aus dem Projekt GRETA. *DIE Z. Erwachsenenbildung* **2016**, *4*, 40–41.
40. Lencer, S.; Strauch, A. Das GRETA-Kompetenzmodell für Lehrende in der Erwachsenen-und Weiterbildung. Available online: https://www.die-bonn.de/doks/2016-erwachsenenbildung-02.pdf (accessed on 20 March 2016).
41. Konert, J.; Buchem, I.; Stoye, J. Digitales Kompetenzverzeichnis mit Technologien des Semantic Web. In Proceedings of the DELFI Workshops 2019, Berlin, Germany, 16–19 September 2019; Gesellschaft für Informatik eVz: Bonn, Germany, 2019.
42. Rentzsch, R.; Staneva, M. Skills-Matching und Skills Intelligence durch kuratierte und datengetriebene Ontologien. In Proceedings of the DELFI Workshops 2020, Online, 14–15 September 2020; Gesellschaft für Informatik eVz: Bonn, Germany, 2020.
43. Dahlmeyer, M.P. Semantic Competence Modelling–Observations from a Hands-on Study with HyperCMP Knowledge Graphs and Implications for Modelling Strategies and Semantic Editors. In Proceedings of the DELFI Workshops 2020, Online, 14–15 September 2020; Gesellschaft für Informatik eVz: Bonn, Germany, 2020.
44. Hernandez, N.; Hazem, A. PyRATA, Python Rule-based feAture sTructure Analysis. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, 7–12 May 2018.
45. *Schlüsselverzeichnis für die Angaben zur Tätigkeit in den Meldungen zur Sozialversicherung–Ausgabe 2010*; Bundesagentur für Arbeit: Nürnberg, Germany, 2017. Available online: https://www.arbeitsagentur.de/datei/schlusselverzeichnis-fur-die-angaben-zur-tatigkeit_ba146811.pdf (accessed on 1 August 2023).