M. Klein, D. Krupka, C. Winter., M. Gergeleit, L. Martin (Hrsg.): INFORMATIK 2024,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2024 2019

Extending the German Labor Market Ontology with Online Data

Diana Martić[1], Andreas Fischer[2], and Jens Dörpinghaus[3]

Abstract: The overall aim of this paper is to increase the comprehensiveness of the German Labour Market Ontology (GLMO). The GLMO provides entities for qualifications, such as occupations and training programs, as well as tools and skills. However, like most knowledge graphs, the GLMO provides only partially complete relationships between entities. This, for instance, affects the mappings of related tools, skills, and qualifications. To enrich the GLMO, publicly available data from the platforms of the Federal Employment Agency are extracted and combined with the GLMO. This integration process has led to the creation of additional entity classes for occupational metadata, including activity fields or activity areas. Moreover, additional links between skills and occupations, and between related qualifications have been established.

Keywords: Labor Market research, Ontology Completion, Web Mining

1 Introduction

Knowledge graphs and ontologies for the labor market are powerful tools for predicting and modeling labor demands, analyzing occupational roles and forecasting unemployment rates [Dö23]. These resources support policy makers, education providers, employers, and career counselors in making decisions related to the labor market ecosystem. The European labor market ontology ESCO, for instance, has been employed by the European Centre for the Development of Vocational Training (Cedefop) for real-time, big data analysis of online job vacancies [Jo17]. Labor market knowledge graphs can additionally support job recommendations, where there is little user-item interaction [GKH21]. Knowledge-based recommender systems automate job matching processes by transforming the vacancies and the candidates" qualifications into standardized sets of skills and competencies. This reduces costs and increases the accessibility. In addition, matches between candidate and vacancies can be explained transparently as knowledge is explicitly modeled [DVP15; Jo17].

Despite their potential benefits, this kind of recommendation systems for Continuing Vocational Education (CVET) programs is still rare [Re22]. Currently, search platforms for continuing education programs, such as the one developed by Germany's Federal Employment Agency - in German: "Bundesagentur für Arbeit" (BA) - only allow for keyword filtering. Some CVET providers are already beginning to integrate complex forms

[1] University of Koblenz, Germany, dmartic@uni-koblenz.de
[2] Forschungsinstitut Betriebliche Bildung (F-BB), 90408 Nürnberg, Germany, andreas.fischer@f-bb.de, https://orcid.org/0009-0006-0748-6076
[3] Federal Institute for Vocational Education and Training (BIBB), Germany, University of Koblenz, Germany, doerpinghaus@uni-koblenz.de, https://orcid.org/0000-0003-0245-7752

doi:10.18420/inf2024_175

of semantic search based on Large Language Models into their platforms (e.g., [Fi23; FLP24]). In their future scenario for CVET recommendation systems, Schleiß et al. argue that an ontology is an essential building block that is still missing from these systems [Sc23]. The ontology should contain information about competence goals and related CVET programs. This enables the deduction of learning objectives, based on a user's qualification profile and, in the next step, the matching to the appropriate CVET courses [Sc23]. Sources of knowledge for the development of such an ontology include the recently developed German Labor Market Ontology (GLMO) [Dö23], publicly available sources from the BA such as the Classification of Occupations 2010 version 2020, in German: "Klassifikation der Berufe 2010 Fassung 2020" (KldB), or the CVET database KURSNET [FD24]. Although these sources are readily available, no comprehensive work has been done to integrate them. As a result, the knowledge graph constructed from these sources may be incomplete, limiting the scope of the resulting analyses. Furthermore, the everchanging nature of the labor market leads to a constant adaptation of competence requirements [Šk22]. Therefore, a labor market and CVET knowledge graph needs to be updated regularly with the latest information. Thus, for a knowledge graph that reflects the current state of the labor market, flexible knowledge completion methods are required.

2 Related Work

For automated analysis of labor market data, the situation in German-speaking countries (Germany, Austria, and Switzerland) is not much different from that in English-speaking countries: "Catalogs play a valuable role in providing a standardized language for the activities people perform in the labor market." [Os18] While these catalogs are widely used to create and compute static values, manage labor market and educational needs, or recommend training and jobs, there is no single ground truth [FD24]. According to [RFE21], one reason for this could be the fact that labor market concepts are modeled by multiple disciplines, each with a different perspective on the labor market. While mapping between different standards like the European ESCO and the American O*NET is discussed, see [Gu22], there are only limited mapping approaches between standards so far. This is the first gap: While there is a diverse field of different taxonomies, catalogs, and even word lists used in different institutions and for different research questions, existing tools tend to focus on only one of these perspectives, making more general solutions difficult.

However, the need for more generic methods has been recognized quite early, for example in education, see [Sz06]. Ontologies and ontology-based methods have been widely used, for example for prediction and modeling of workshops and labor market needs, see [BK19], for identification of job knowledge, see [Kh15], but also for analysis of particular jobs, see [Pa19], or for matching educational content to generic texts, see [Po21]. They have also been used to predict the unemployment rate, see [Li14]. However, these approaches mainly focused on a particular labor market characteristic, such as skills, knowledge, educational content, or job classifications.

The most general approach that exists so far is called OntoJob, where natural language processing is used to train a labor market ontology from job advertisement texts, see [Vr22]. However, this approach has several serious limitations, in particular it is not linked to established de-facto standards such as ESCO. This was the reason for the development of the German Labor Market Ontology (GLMO) [Dö23] which was used in several research projects [Dö22; DSH23; DT23; DWB22; FD24].

The aim of this work is to enrich the GLMO. This ontology, developed by the BIBB, was primarily created based on the multilingual, European ontology for occupations and skills ESCO and the German classification of occupations KldB. In addition to the full hierarchy of KldB occupations, the GLMO contains sets of skills, tools, and educational trainings relevant to the German labor market which are defined by the BA and will be denoted as BA skills, tools, and educational trainings in the following. In the ontology, the concepts are hierarchically organized and mapped to KldB-occupational unit groups [Dö23]. However, the mappings do only cover a subset of nodes. Additionally, information provide by publicly accessible BA resources, which will be discussed in the next section, have not been integrated.

3 Method

The following section introduces the GLMO and new data sources on the German labor market that are not yet fully integrated into the GLMO.

3.1 GLMO

Dörpinghaus et al. have developed a German labor market ontology that combines the BA's KldB, European Skills, Competences, Qualifications and Occupations (ESCO), ISCO-08 and several currently unpublished taxonomies and mappings [Dö23]. Thereby, the GLMO includes the entire KldB hierarchy and historical occupational data, where each instance is described by the occupation's title[4]. Additionally, for all 3-digit KldB-codes, it offers mapping to WZ08 industrial sectors. For the 5-digit KldB-codes, 43,452 relations to a subset of the 9,078 BA skills and 10,991 BIBB tools are included [Dö23]. In contrast, ESCO does not make a clear distinction between skills and tools.

The GLMO establishes mappings from each KldB unit group to an ISCO unit group. Since ESCO also maps all its occupations to ISCO groups, this allows for loose links between KldB unit groups and ESCO occupations. In addition, the GLMO offers mappings for 3,054 BA skills to a subset of the 13,485 ESCO skills [Dö23].

[4] Extracting further details from the KldB such as common activities or the full descriptions is an open challenge, that will not be addressed in this paper.

In contrast to ESCO, the GLMO includes concepts to model educational programs. The GLMO knowledge graph has since been revised by the BIBB to also include nodes for educational trainings, specifically `glmo:ContinuingVocationalEducation`, `glmo:GeneralEducation`, `glmo:ContinuingEducation`, and `glmo:Apprenticeship`. For the instances of the latter two concepts, partial mappings to occupations exist. Relations to skills and tools have not been incorporated. Using the SKOS ontology, the GLMO constructs hierarchical relationships with `skos:broader`. These hierarchies are available for occupation (ISCO and KldB), skill, tool, industrial sector, general education, and partially for continuing vocational education instances.

3.2 Data

The selected resources are all administered by the BA and consist of the platforms BERUFENET, New Plan and KURSNET. For each platform, the BA provides API endpoints that have been documented by the civil society initiative bund.dev [FD24].

To complement this knowledge graph, the BA platforms KURSNET, BERUFENET and New Plan are promising resources:

- KURSNET is the BA's database and online portal for vocational training and further education in Germany (cf. [FD24]). Among other things, it provides information on two types of CVET programs, namely those for regulated qualifications and those that do not formally qualify for another occupation. Offers of the former type are linked to the KldB position they qualify for. The majority of these categories has already been modeled in the GLMO. However, neither KURSNET nor the GLMO provide direct mappings to the BA skills that are targeted by a CVET category.
- BERUFENET is a database and online portal that provides informational material on KldB occupational titles and groups them into study fields, activity fields, and activity areas. This also includes mappings to related/alternative occupations, further qualifying and other CVET categories from KURSNET as well as info fields with extensive additional information (e.g., competencies, abilities, knowledge and skills, cf. [FD24]). However, the platform only covers a subset of 3,574 of the total 31,610 KldB occupation titles listed in the DKZ and none of the KldB groups (one to five digits).
- New Plan is a database and online portal for professional reorientation (cf. [FD24]) that provides an overview over alternative occupations, which are proposed based on occupation change statistics. For the selected occupations, New Plan provides lists of associated BA skills, distinguishing between optional and essential skill. This skill set includes 13,000 skills, apart from those already present in the GLMO. However, the mapping is only provided for approximately 4,000 occupations.

3.3 Extension Method

First, ESCO includes nodes for ISCO, as each occupation is mapped to an ISCO unit group. These nodes were connected to the `glmo:ISCO` nodes through a `skos:closeMatch` relation using a mapping file provided by the BA. Then, mappings between ESCO and BA skills were established through skos:closeMatch relations, using the mapping files provided by the BIBB.

The further education opportunities, other CVETs and alternative occupations listed in BERUFENET and KURSNET were mapped to the GLMO occupation or CVET nodes, respectively, through the systematic position or the DKZ ID. The DKZ ID is a technically generated identification number for each individual data record in the DKZ.

Likewise, the alternative unit groups, suggested 8-digit occupations, and skills from New Plan were mapped via the DKZ ID, or in the case of skills, via the systematic position.

To represent the additional information provided by the BA data sources, the schema of the GLMO was extended. The modifications include the introduction of the concept `glmo:OccupationalMetaData` and its subclasses. `glmo:IndustrialSector` is a concept already present in the original GLMO, that was modeled as a subclass. Furthermore, to incorporate the BERUFENET categorization systems, the classes glmo:StudyField, `glmo:ActivityField`, and `glmo:ActivityArea` were introduced. The `skos:related` relation can be employed to establish connections between occupations and occupational metadata.

Additionally, a concept hierarchy for qualifications was introduced, with `glmo:Qualification` serving as the most abstract concept, distinguishing between `glmo:Education` and `glmo:Occupation`. Instances of `glmo:Qualification` can be connected via `glmo:isQualificationFor`. This accommodates scenarios where not only educational programs can be qualifications for occupations, but also occupations qualify for educational programs.

To model the suggestions made by New Plan and alternatives from BERUFENET, the relation type `glmo:hasAlternative` was added. This relation connects two `glmo:Occupation` instances, representing alternatives for a specific occupation. For example, alternatives for the occupation “Data science, data science (undergraduate)” include “Market researcher” or “Data engineer.” Additionally, the relations `glmo:hasSubjectArea` and `glmo:hasAdvancedTraining` can be employed for meta subject areas and advanced training from BERUFENET. The occupation (8-digit) “livestock farmer - cattle farming”, for instance, has the subject area “livestock farmer.” These three relations have been added instead of using `skos:related` to preserve the narrower semantic context of the relation.

Then, for distinguishing between qualifications leading to an occupation, qualifications qualifying for an educational program, and additional CVETs that are more general, the relation glmo:hasFurtherTraining is introduced. For example, “interior design” has the further training “construction project management,” but participating in this CVET does

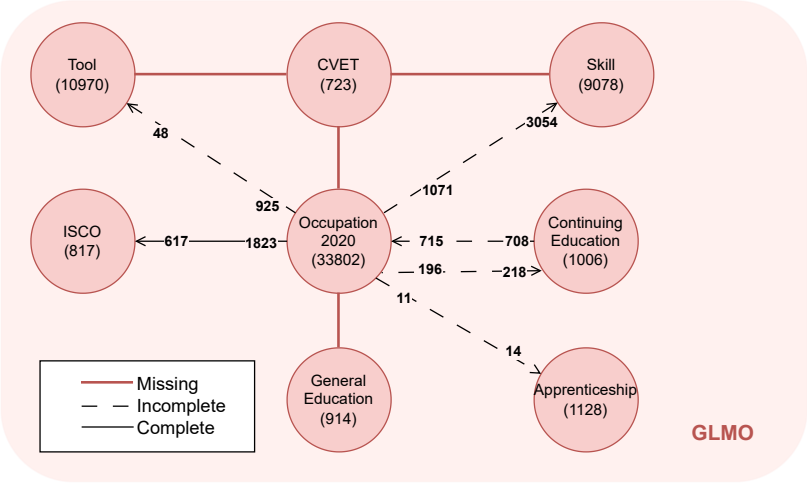

Figure 1: Overview over missing (red line) and incomplete (dashed line) relationships in the initial version of the GLMO

not directly qualify for another occupation. To link GLMO skills and ISCO groups with ESCO skills, the `skos:closeMatch` or `skos:exactMatch` relations can be used.

4 Evaluation

Here, we will outline acquisition and preprocessing steps for the data sources and the modifications of the GLMO are described. We will also discuss the completeness of the resulting knowledge graph.

Figure 1 gives an overview over the missing or incomplete mappings between the entities in the initial version of the GLMO (excluding occupational metadata). There, the number in brackets indicates how many nodes of this type exist in total. If the mapping between two types of nodes is incomplete, i.e. if they are connected through a dashed line, then the edge contains two counts, indicating the number of nodes of the involved types. For the edge between “Occupation2020” and “skill” this means, that in all relationships between these two types of nodes, there are 1071 unique occupations and 3054 unique skills involved. These results were obtained through the following query:

```
MATCH (s:Skill)<-[]-(o:GLMO:Occupation) RETURN COUNT(DISTINCT s), COUNT(DISTINCT o).
```

4.1 BERUFENET

For sourcing BERUFENET, initially the list of all DKZ IDs was requested. From this, 3,566 KldB-codes were retrieved. Using these DKZ IDs, information on suitable trainings, thereby differentiating between further trainings (“aufstiegsweiterbildungen”), which are occupations, and other CVETs (“anpassungsweiterbildungen”) were obtained [FD24].These fields are all encoded as JSON objects, that include the respective DKZ-IDs of the trainings and occupations[5].

BERUFENET moreover categorizes the occupations by field of study, field of activity and area of activity. To obtain the activity area list, endpoint 4 was queried starting with the systematic position “TF”. The resulting list was used to extract the systematic positions of lower-level activity areas, which was then iteratively queried again. Similarly, study areas were extracted taking systematic position “HA” as a starting point. To create mappings to occupations, the occupation DKZ ID was used to request for activity areas, study fields activity fields.

4.2 New Plan

The list of BA skills was retrieved from the New Plan API [FD24]. These skills include all skills already present in the GLMO and are described by the DKZ ID, their systematic position, the systematic position of the higher-level skill and the preferred label. Using KldB-codes, skill mappings were obtained. The returned values contained the DKZ IDs of the skills and a marker for whether it is a mandatory or an optional skill for the queried occupation. Rather than accessing the BERUFENET API for the skills mappings, we opted for the New Plan API, as it provides the mappings in JSON objects. This enabled the extraction process to proceed without the need for additional steps to cleanse and parse the HTML fragments. Furthermore, it includes additional codes (“codenr” and “obercodenr”) that model the particular skills’ position in the skill hierarchy. These codes are not available at the BERUFENET APIT.

For the suggestions, the API was requested with the DKZ ID of a 5-digit KldB-code. The response contains mappings to the DKZ ID of related 5-digit KldB-codes along the proportion of switches based on the associated 3-digit KldB-codes. For each requested DKZ-ID, there is a distinction made between near switches, where the leading 3-digits of the KldB systematic position matches, and far switches. Additionally, the API provides a list of DKZ IDs of a selection of 8-digit KldB-code for a given 5-digit KldB-code. In most, but not all, cases, the leading five digits of the suggestions matched with the queried 5-digit KldB-codes.

[5] The API provides further details on tasks and activities, work conditions, personal criteria, interests, prospects, and trends. However, this data is semi-structured in HTML fragments and not integrated into the enriched KG. Its processing remains an open task.

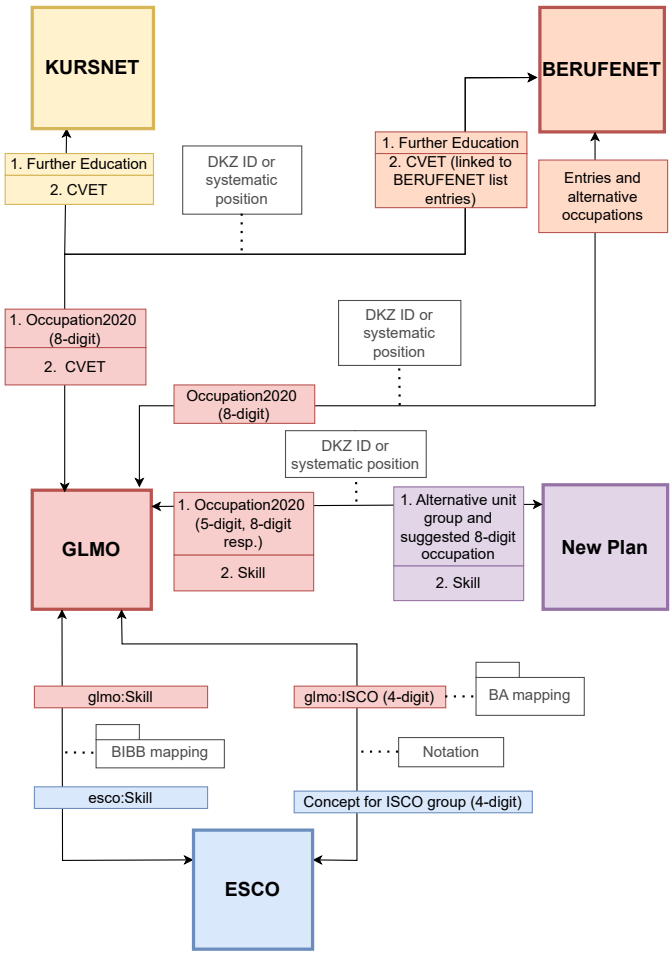

Figure 2: Connections and mappings between ESCO, the GLMO and the BA platforms

4.3 KURSNET

The KURSNET API was requested to obtain the full set of keywords for a CVET as defined by the BA via its DKZ ID [FD24]. The initial set of DKZ IDs was provided by the BIBB. Additional CVETs were retrieved by querying the API with potential DKZ IDs between those already known. Finally, the number of offers per CVET category was requested.

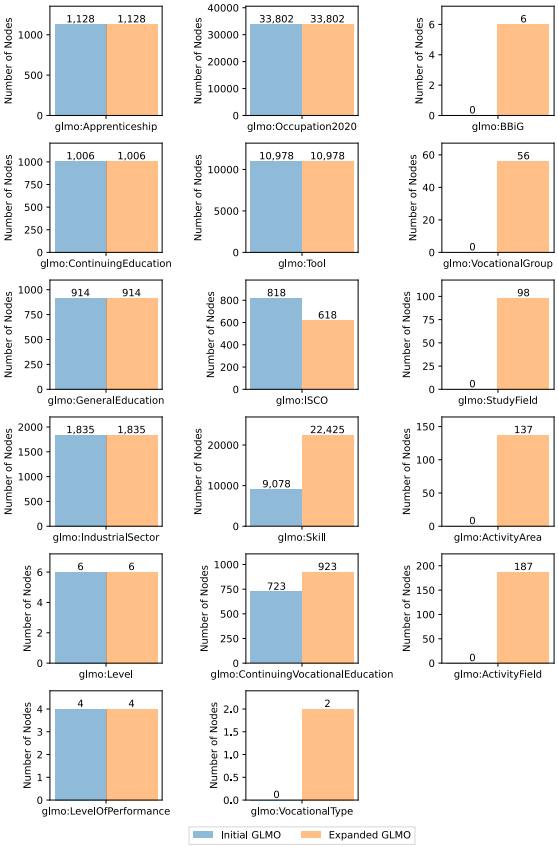

Figure 3: Comparison of instances in the initial GLMO and the revised GLMO. The blue bars shows the number of instances per node type for the initial GLMO. The orange bar shows the number of node instances in the revised GLMO.

4.4 Integration Results

Figure 2 describes how the existing GLMO nodes can be combined with the additional data sources. The number of nodes per label resulting from the integration process is displayed in Figure 3. No changes were made to the tool node set. Most of the qualification subclasses did not have different numbers of nodes, as the original GLMO was already highly comprehensive in this dimension. The changes in the number of ISCO nodes were due to the cleanup process, as described earlier. For the education subclass, only the number of CVETs increased. Additional CVETs in KURSNET are not linked to any specific offer and are instead used primarily to establish a hierarchy and group CVETs. However, a partial fill-in of the missing mapping between CVET and occupations was achieved with 26,809

new relationships, involving 4,809 of the 31,610 8-digit occupations and 613 CVETs from the set of CVETs that were already present in the original GLMO.

The comparison of the number of relationships for each label and relation combination was evaluated. For all remaining education sub-classes, the gap between occupations and skills mappings remains. However, 4,722 8-digit KldB-codes have additional training options and 4,106 occupations have alternatives. New nodes were included for occupational metadata and particularly for skills. In the original GLMO there were 6 levels in the hierarchy of skills. Comparing to this, in the skills list obtained from New Plan a new level was added at the bottom of the hierarchy. Out of the 13,347 new skills, 12,107 lie in this new bottom level. However, these skills were not associated with any occupations, resulting in a new gap in the skill-occupation mapping. Of the 1,240 new skills from the existing hierarchy levels, only 35 have been mapped to occupations. These mappings involve 657 8- digit KldB-codes in 793 relations. The majority of the new relations between occupations and skills involve skills that were already present in the initial GLMO. In the initial GLMO, there are 13,802 relations involving 3,054 skills and 1,071 5-digit KldB-codes (refer to Figure 1). With the new mappings, 173,556 relations were added, involving 6,109 8-digit occupations and 5,379 skills that were already present in the original GLMO. Out of these skills, 2,325 were not mapped in the initial GLMO, leaving 6,024 previously listed skills still unmapped. The new occupational metadata subclasses for activity and study field only cover a portion of the KldB-codes. The activity field mappings include 4,989 8-digit occupations and the study field mappings includes 362 8-digit-codes. The difference arises because only some occupations require a university degree. Most occupations are not covered. However, it was possible to fully construct the hierarchies of the activity fields, activity areas and study fields.

5 Conclusions and Outlook

In this paper, the GLMO has been enriched using publicly available BA data. Occupational metadata, including activity fields and alternatives, were added through this process. In addition, this integration enabled the completion of the CVET hierarchy, the addition of fine-grained skills and the addition of relationships from fine-grained occupations to skills, CVET or alternatives and suggestions. The inclusion of the new skills, however, increased the mapping gap of the initial GLMO. The mapping gaps for educational training, skills and tools remained unchanged. Additionally, none of the previously identified mapping gaps could be entirely resolved.

One limitation of this work is that the BA endpoints accessed offer additional sources of knowledge, such as detailed task descriptions for occupations, which are encoded in HTML fragments. These fragments were not extracted in this paper.

The BA classifications are used for the annotation of specific offers, such as further education offers from KURSNET, job offers and applicant profiles. As the GLMO maps

these classifications, it offers a wide range of possible analyses of the labor market and its requirements. These analyses might include analyses on how the skills mentioned in job offer differ from the skills mentioned by candidates. For potential competency gaps, the number of related CVET offers could be analyzed.

References

[BK19] Boldyreva, E.; Kholoshnia, V.: Ontological Approach to Modeling the Current Labor Market Needs for Automated Workshop Control in Higher Education. In: MICSECS. 2019.

[Dö22] Dörpinghaus, J. et al.: From social networks to knowledge graphs: A plea for interdisciplinary approaches. Social Sciences & Humanities Open 6/1, p. 100337, 2022.

[Dö23] Dörpinghaus, J. et al.: Towards a German labor market ontology: Challenges and applications. Applied Ontology/18(4), pp. 1–23, 2023.

[DSH23] Dörpinghaus, J.; Samray, D.; Helmrich, R.: Challenges of Automated Identification of Access to Education and Training in Germany. Information 14/10, p. 524, 2023.

[DT23] Dörpinghaus, J.; Tiemann, M.: Vocational Education and Training Data in Twitter: Making German Twitter Data Interoperable. Proceedings of the Association for Information Science and Technology 60/1, pp. 946–948, 2023.

[DVP15] De Smedt, J.; le Vrang, M.; Papantoniou, A.: ESCO: Towards a Semantic Web for the European Labor Market. In: Ldow@ www. 2015.

[DWB22] Dörpinghaus, J.; Weil, V.; Binnewitt, J.: Towards the Analysis of Longitudinal Data in Knowledge Graphs on Job Ads. In: The Workshop on Computational Optimization. Springer, pp. 52–70, 2022.

[FD24] Fischer, A.; Dörpinghaus, J.: Web Mining of Online Resources for German Labor Market Research and Education: Finding the Ground Truth? Knowledge 4/1, pp. 51–67, 2024.

[Fi23] Fischer, A. et al.: KI-basierte Personalisierung berufsbezogener Weiterbildung. In: Leitfaden für die Bildungspraxis, Band 73. wbv, pp. 1–40, 2023.

[FLP24] Fischer, A.; Lorenz, S.; Pabst, C.: Empfehlungen zur beruflichen Weiterbildung - Entwicklung eines KI-basierten Entscheidungsmanagements. Berufsbildung in Wissenschaft und Praxis 1/2024/53, pp. 32–34, 2024.

[GKH21] Guruge, D. B.; Kadel, R.; Halder, S. J.: The state of the art in methodologies of course recommender systems—a review of recent research. Data 6/2, p. 18, 2021.

[Gu22] Guru Rao, S.: Ontology Matching using Domain-specific knowledge and Semantic Similarity, MA thesis, University of Twente, 2022.

[Jo17] Jones, D.: ESCO handbook: European skills, competences, qualifications and occupations. Publication Office of the European Union, Luxembourg, 2017.

[Kh15] Khobreh, M. et al.: An ontology-based approach for the semantic representation of job knowledge. IEEE Transactions on Emerging Topics in Computing 4/3, pp. 462–473, 2015.

[Li14] Li, Z. et al.: An ontology-based Web mining method for unemployment rate prediction. Decision Support Systems 66/, pp. 114–122, 2014.

[Os18] Ospino, C.: Occupations: Labor Market Classifications, Taxonomies, and Ontologies in the 21st Century. Inter-American Development Bank/, 2018.

[Pa19] Papoutsoglou, M. et al.: Extracting knowledge from on-line sources for software engineering labor market: A mapping study. IEEE Access 7/, pp. 157595–157613, 2019.

[Po21] Poletaikin, A. et al.: Ontology Approach for the Intelligent Analysis of Labor Market and Educational Content Matching. In: 2021 International Symposium on Knowledge, Ontology, and Theory (KNOTH). IEEE, pp. 50–55, 2021.

[Re22] Reichow, I. et al.: Recommendersysteme in der beruflichen Weiterbildung. Grundlagen, Herausforderungen und Handlungsempfehlungen. Ein Dossier im Rahmen des INVITE-Wettbewerbs, 2022.

[RFE21] Rodrigues, M.; Fernández-Macías; Enrique, Sostero, Matteo: A unified conceptual framework of tasks, skills and competences, ed. by European Commission, Seville, 2021.

[Sc23] Schleiß, J. et al.: Künstliche Intelligenz in der Bildung. Drei Zukunftsszenarien und fünf Handlungsfelder. KI-Campus/, 2023.

[Šk22] Škrinjarić, B.: Competence-based approaches in organizational and individual context. Humanities and social sciences communications 9/1, pp. 1–12, 2022.

[Sz06] Szabó, I.: The implementation of the educational ontology. In: Proceedings of the 7th European Conference on Knowledge Management, Corvinus University of Budapest, Hungary, ACL, UK. Pp. 541–547, 2006.

[Vr22] Vrolijk, J. et al.: OntoJob: Automated Ontology Learning from Labor Market Data. In: 2022 IEEE 16th International Conference on Semantic Computing (ICSC). IEEE, pp. 195–200, 2022.