universität
koblenz
weiter:denken

Master Thesis Exposé

Building an efficient approach towards mapping, cross-domain enrichment of dual-lingual ontologies

Submitted By

Venkatesh Hariharapura Shivashankar [220 200 713]

First Supervisor: Prof. Dr. Frank Hopfgartner
Institute for Web Science and Technologie(WeST)
University of Koblenz

Second Supervisor: Dr. Jens Dörpinghaus
University of Koblenz
and
Federal Institute for Vocational Education and Training (BIBB),
Bonn

Koblenz, May 2023

1 Introduction

1.1 Background and Context for the work

With the increasing globalization of the internet, there is a growing need for processing data in multiple languages. This has led to a significant increase in the amount of multilingual data available on the web, which presents a major challenge for accessing, processing, and integrating data from different language sources [3]. An Ontology provides a shared and precise source for system interoperability and reuse of knowledge base [18]. In this study, the main focus will be on how we can efficiently build an approach which maps and enriches cross domain dual lingual ontology.

It is crucial for individuals, companies, and governments to comprehend developments in the labor market and how tools associated to occupations have changed over time. Up to 375 million workers may need to shift occupational categories and pick up new skills by 2030 as a result of automation and other technological advancements, according to a McKinsey Global Institute report [12]. Also, job definitions are continually changing as new jobs are created and old ones are reinterpreted. According to a World Economic Forum report [6], workers must retrain and upgrade their skills to stay up with the shifting needs of the labor market. Furthermore, recognizing new employment opportunities and the abilities essential to succeed in them requires a grasp of how tools associated to jobs have changed. People and organizations who can adopt new tools and skills will be better positioned for success as technology continues to evolve and alter the employment market.

In this thesis, a legacy tools taxonomy generated in 2018 through the analysis of the Bundesinstitut für Berufsbildung (BIBB)-Bundesanstalt für Arbeitsschutz und Arbeitsmedizin (BauA) employment survey [9], German job advertisements data, and the Computer Science Ontology (CSO, Version: v.3.2) will serve as the primary source of data. To enhance the existing taxonomy, I intend to utilize the CSO and dual lingual translation to incorporate additional information. The resulting taxonomy will be analyzed using the latest German job advertisement dataset to identify trends in the development and/or requirement of tools in the Computer Science domain.

The data from job advertisements along with the enriched tools taxonomy will be used to carry out empirical study. While job advertisements may not focus largely on the tools used on a daily basis, they are nevertheless helpful for learning about new technologies or pertinent tools. So, it is possible to trace the adoption of digital work equipment in the workplace and changes in professional activities over time by looking at job advertisements and the changes in the tools taxonomy. Other information included in job advertisements can further shed light on the relationships between tools of the trade, educational requirements, professions, and industry sectors. In general, job advertisements are a useful source of data for researching shifts in the labor market.

1.2 Definitions of specialist terms

Computer Science Ontology (CSO)[1] is a comprehensive ontology of research areas, topics, and concepts in the field of Computer Science. It provides a formal description of the research areas, topics, and concepts within the discipline of Computer Science, which can be used to facilitate the understanding and discovery of research content.

GLMO[2] is a German labor market ontology that builds upon European Skills, Competences, Qualifications and Occupations (ESCO)[3] and Simple Knowledge Organization System (SKOS)[4] to integrate International Standard Classification of Occupations (ISCO)[5] and Klassifikation der Berufe (KldB)[6], with a focus on occupations, skills, tools, and industrial sectors. GLMO provides a standardized format for mappings and additional data that are not yet available in ESCO.

1.3 Aim and Objectives

The aim of this study is to develop an efficient approach towards mapping, cross-domain enrichment and quality control of dual-lingual ontologies.

Objectives: The main objective of this work is to develop an efficient approach towards mapping, cross-domain enrichment, and quality control of dual-lingual ontologies. To achieve this, several sub-objectives have been identified. Firstly, the current methodologies for mapping multi-lingual ontologies will be analyzed to identify the challenges in

[1] https://cso.kmi.open.ac.uk/home

[2] https://tm4vetr.github.io/glmo/

[3] https://esco.ec.europa.eu/de

[4] https://www.w3.org/2004/02/skos/

[5] https://www.ilo.org/public/english/bureau/stat/isco/

[6] https://statistik.arbeitsagentur.de/DE/Navigation/Grundlagen/Klassifikationen/Klassifikation-der-Berufe/Klassifikation-der-Berufe-Nav.html

the existing methods. Secondly, the gaps in legacy tools taxonomy and CSO will be identified to understand how to enrich tools taxonomy as part of GLMO using CSO. Thirdly, the potential of job advertisements and an enriched tools taxonomy in the domain of computer science will be explored to track the evolution of technological adoption and changes over time. Fourthly, additional insights will be obtained regarding the relationships between tools and job advertisements. Finally, the proof of concept will be provided using real-world examples to demonstrate the practical implementation of the proposed approach. By achieving these objectives, the study will contribute to the development of a comprehensive approach for efficient mapping, cross-domain enrichment, and quality control of dual-lingual ontologies.

2 Literature review

2.1 Literature Search Strategy

To conduct a comprehensive literature search on the topic of building an efficient approach towards mapping, cross-domain enrichment, and quality control of dual-lingual ontologies, a search strategy using relevant keywords such as "CSO," "ontology engineering," "dual-lingual ontology enrichment," "BIBB," and "tools taxonomy" should be employed. This search should include popular databases such as Google Scholar, ACM Digital Library, IEEE Xplore, and ScienceDirect. The search should focus on identifying existing methodologies for mapping multilingual ontologies, gaps in legacy tools taxonomy and CSO, techniques for enriching tools taxonomy in GLMO using CSO, and examples of how job advertisements and an enriched tools taxonomy in the domain of computer science can be used to track the evolution of technological adoption and changes over time.

2.2 The Literature Review

CSO is vast in nature and was generated using 16 million publications which are related to the field of Computer Science [17]. To improve the accuracy of scientific bibliographic databases and researchers' ability to identify pertinent research publications, the CSO was established [14].

Gaps in an ontology refer to areas or concepts inside a domain that the ontology does not describe or cover. According to Gangemi and Presutti [16], gaps in the ontology can cause problems like lack of interoperability with other ontologies, difficulty in information retrieval, and incomplete or inaccurate representation of the domain.

The concept of work equipment has been evolving in response to technological progress and the changing nature of work. Work equipment encompasses both material and immaterial objects that are necessary for professional activities. The use of work equipment requires specific skills and abilities, which are closely linked to job descriptions and the work to be performed [7]. Systematically categorizing work equipment can provide insights into changes in professional activities and requirements over time. To this end, BIBB developed a taxonomy for work equipment that is comprehensive. This scheme is subject to ongoing updates and can be applied to different data sets and questions, making it a valuable tool for analyzing changes in the job market [7].

Dual lingual ontology enrichment is a well-studied topic in the field of ontology engineering and natural language processing. Numerous researchers have proposed different approaches and methods for automatically enriching ontologies with multilingual information. For instance, a study by Ibrahim et al. (2019) proposes a fully automated cross-lingual matching approach for ontology enrichment, which selects the best translation based on semantic similarity for building multilingual ontologies [10]. A study by Bouscarrat et al. (2020) examines the use of open-source knowledge bases like Wikidata for translating biomedical ontologies in multiple languages, focusing on the coverage and quality of the translations [1]. The study analyzes the direct and second-order links between two biomedical ontologies and Wikidata for 9 European languages, Arabic, Chinese, and Russian. It assesses the accuracy of translations by comparing them to those generated by a commercial machine translation tool [1]. In addition to these approaches, several researchers have proposed different systems for dual lingual ontology enrichment. For example, in a study by Espinoza et al. (2008), the authors presented a system called Label-Translator for automatically localizing ontologies in a multilingual setting [5].

Hailu et al. (2014) used DBpedia, a structured version of Wikipedia, to translate Gene Ontology terms from English to German. They performed keyword-based searches in the DBpedia dataset and selected translations based on the presence of corresponding English and German Wikipedia pages. Approximately 25% of the terms had translations available in both languages, using the keyword-based Wikipedia URI lookup [8].

Since the CSO contains comprehensive information about tools, products, software, and hardware related to each entity, it serves as a valuable source for extracting necessary information. However, the existing legacy tools taxonomy is outdated and has significant gaps, particularly in terms of incorporating the latest tools in the computer science domain. Therefore, it is crucial to update and enrich the taxonomy with current tools.

The GLMO, developed by BIBB, is an ontology that encompasses occupations, skills, tools and industrial sectors. Currently, the tool section of the GLMO is empty, and our objective is to populate it by integrating the tools from the CSO. Since the CSO is primarily in English, and the GLMO is predominantly in German, the need for dual lingual translation arises to bridge the language barrier and enable seamless integration.

Although the studies by Hailu et al. (2014) [8] and Bous-carrat et al. (2020) [1] focused on extracting translation for dual lingual ontologies using DBpedia and Wikipedia, there is currently a research gap in extracting tools related to each topic from both DBpedia and Wikipedia sources. To address this gap, I propose a cross-domain dual lingual translation and ontology enrichment approach. By leveraging this approach, we aim to enhance the GLMO with a comprehensive list of tools by linking them to their respective topics. This will facilitate the classification of every occupation with its related tools using the "isPartOf" relation in the GLMO.

To achieve this, we will utilize the language tags available in the web content of each link in Wikipedia and DBpedia. These language tags provide information about the language of the content, allowing us to retrieve the corresponding names of the links in any desired language. Specifically, we will focus on obtaining the English and German names of each tool. By utilizing the language tags and web content, we can establish a foundation for the dual lingual translation between the English and German names of each tool.

This methodology will enable us to bridge the language barrier between the CSO and the GLMO. By incorporating dual lingual translation, we can enrich the GLMO with a vast collection of tools from the CSO.

In summary, the goal of this thesis is to provide a taxonomy of tools related to entities in CSO and a Named Entity Recognition (NER) list containing topics and their corresponding tool lists. This NER list will be a valuable resource for future research and practical applications, enabling the classification and analysis of occupations based on their associated tools within the context of the enriched GLMO.

2.3 The Research Question

Main research question of this work is: “How can we build an efficient approach towards mapping, cross-domain enrichment of dual-lingual ontologies?”

Sub research questions that I would answer through my study are as below:

- What are the current methodologies for mapping multi-ligual ontologies and what are the challenges in the existing methods?

- What are the gaps in legacy tools taxonomy and CSO?
- How to enrich tools taxonomy using CSO?
- How can job advertisements and an enriched tools taxonomy in the domain of computer science be utilized to track the evolution of technological adoption and changes over time? What additional insights can be obtained regarding the relationships between tools and job advertisements?
- How to provide proof of concept using real world examples?

3 Methodology planning

3.1 Data Collection Procedure

In this study, I will utilize the 2018 legacy tools taxonomy available at BIBB, along with the v.3.2 version of CSO, which will be downloaded from the official website in a comma-separated values format. Moreover, I will also collect the most recent job advertisement data from BIBB for our analysis.

3.2 Planned Methodology

Given the nature of my study, I have opted for an inductive approach, which involves the initial collection of data on tools within each topic of CSO and the subsequent translation of tool names in dual languages. The inductive approach aligns well with my research objectives of enriching GLMO with comprehensive tool information and creating a Named Entity Recognition (NER) list.

Quantitative research methodology is a systematic approach to collect and analyze numerical data to uncover patterns, relationships, and trends [19]. On the contrary, qualitative research methodology centers around the collection and analysis of non-numerical data, aiming to understand and interpret subjective experiences, meanings, and perspectives [19].

Considering the nature of tools, especially in the field of computer science, can be diverse and evolving, making it challenging to assign numerical values or categories to capture their full range. Hence, quantitative research in this study is not appropriate. By employing qualitative research, a thorough understanding of the tools and their applicability within the domain can be achieved, which can offer insightful data that may not be clearly measurable.

In the context of fetching tools from CSO and enriching the ontology, a qualitative research methodology is better suited for many reasons. Firstly, qualitative research allows for a comprehensive exploration of the tools related to all the entities in CSO. Secondly, qualitative research enables the identification of emerging patterns, themes, and trends within the collected data. This qualitative exploration allows for a better contextual understanding of the tools, which can inform the enrichment of the ontology and the creation of a comprehensive NER list. Furthermore, as the field of computer science constantly evolves with new tools and technologies, qualitative methods allow me to stay current and capture the latest developments.

The effectiveness of the dual lingual ontology mappings approach is evaluated using evaluation measures including precision, recall, and F-measure. These quantitative measurements give clear indications of the precision, comprehensiveness, and general effectiveness of the mappings. By calculating the percentage of relevant mappings among the retrieved mappings, precision assesses the accuracy of the mappings. By calculating the percentage of relevant mappings that were successfully retrieved, recall quantifies how thorough the mappings are. The F-measure combines recall and precision to offer an objective evaluation of the mappings' quality. By using these assessment measures, the effectiveness of the strategy may be rigorously and thoroughly evaluated, allowing for data-driven insights and wise decision-making.

The methodology for this study consists of two main steps, with the possibility of developing an automated methodology for future analyses:

Step 1: Update GLMO using current CSO: The current CSO and tools taxonomy will be used to update the GLMO. The update will be conducted through a dual-lingual translation process to ensure uniformity and consistency. The output of this process will be an enriched tools taxonomy in GLMO and CSV word list for Named Entity Recognition (NER).

Step 2: Analyze changes in the job market using job advertisement data using qualitative methods: The job advertisement data from BIBB[1] will be analyzed along with the legacy tools taxonomy, the latest job advertisement, and the enriched tools taxonomy. This analysis will enable the identification of changes in the job market, specifically in terms of the trends in requirements and tools used, particularly in the computer science domain. A detailed analysis of these trends will be provided.

[1] https://www.bibb.de/

An automated methodology will be developed to streamline the entire data pipeline and integrated into existing Text Mining for VET Research[2] pipeline, encompassing steps 1 and 2. This automation will enable the enrichment of the tools taxonomy of GLMO and analysis of the job market for each new version of CSO and job advertisement in the future.

3.3 Consideration of Ethical Issues

For my thesis, I will be using several sources of data including the CSO from its official website as well as Legacy Tools Taxonomy, and German job advertisement data provided by BIBB. I am fully committed to considering all ethical issues that may arise during my research process. I understand the importance of ensuring that my research is conducted in a manner that is respectful of human subjects and their rights. Therefore, I will take the necessary steps to protect the confidentiality and privacy of any individuals who are represented in my data. Additionally, I will ensure that my use of these sources of data is in accordance with all relevant laws, regulations, and ethical standards. Overall, I will strive to conduct my research with the highest ethical standards and contribute to the advancement of knowledge in a responsible and ethical manner.

3.4 Project Timetable

[2] https://github.com/TM4VETR/

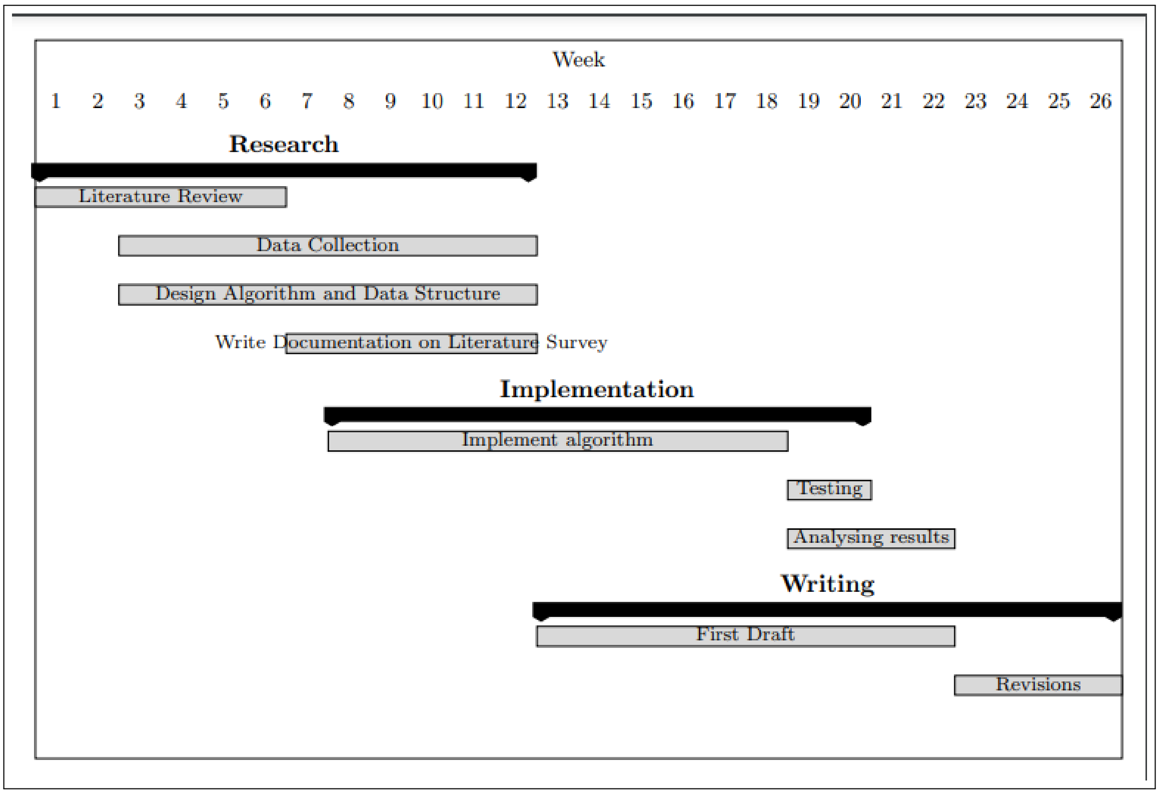

Figure 3.1: Preliminary Schedule

Bibliography

[1] L. Bouscarrat, A. Bonnefoy, C. Capponi, and C. Ramisch. Multilingual enrichment of disease biomedical ontologies. *arXiv preprint arXiv:2004.03181*, 2020.

[2] A. Chauhan, L. Sliman, et al. Ontology matching techniques: a gold standard model. *arXiv preprint arXiv:1811.10191*, 2018.

[3] Y. Dang, Y. Zhang, P. J.-H. Hu, S. A. Brown, Y. Ku, J.-H. Wang, and H. Chen. An integrated framework for analyzing multilingual content in web 2.0 social media. *Decision Support Systems*, 61:126–135, 2014.

[4] J. Dörpinghaus, V. Weil, and J. Binnewitt. Analyzing longitudinal data in knowledge graphs utilizing shrinking pseudo-triangles. In *2022 17th Conference on Computer Science and Intelligence Systems (FedCSIS)*, pages 323–327, 2022.

[5] M. Espinoza, A. Gómez-Pérez, and E. Mena. Enriching an ontology with multilingual information. In S. Bechhofer, M. Hauswirth, J. Hoffmann, and M. Koubarakis, editors, *The Semantic Web: Research and Applications*, pages 333–347, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg.

[6] W. E. Forum. The future of jobs report 2020, 2020.

[7] B. Güntürk-Kuhl, A. C. Lewalder, and P. Martin. Die taxonomie der arbeitsmittel des bibb. *BIBB Fachbeitr"age zur beruflichen Bildung*, (2017), 2017.

[8] N. D. Hailu, K. B. Cohen, and L. E. Hunter. Ontology translation: A case study on translating the gene ontology from english to german. In E. Métais, M. Roche, and M. Teisseire, editors, *Natural Language Processing and Information Systems*, pages 33–38, Cham, 2014. Springer International Publishing.

[9] A. Hall and D. Rohrbach-Schmidt. Bibb/baua employment survey 2018. *BIBB-FDZ - Daten- und Methodenberichte*, (2020), 2020.

[10] S. Ibrahim, S. Fathalla, H. Shariat Yazdi, J. Lehmann, and H. Jabeen. From monolingual to multilingual ontologies: The role of cross-lingual ontology enrichment. In M. Acosta, P. Cudré-Mauroux, M. Maleshkova, T. Pellegrini, H. Sack, and Y. Sure-Vetter, editors, *Semantic Systems. The Power of AI and Knowledge Graphs*, pages 215–230, Cham, 2019. Springer International Publishing.

[11] B. T. Le, R. Dieng-Kuntz, and F. Gandon. On ontology matching problems. *ICEIS (4)*, pages 236–243.

[12] J. Manyika, S. Lund, M. Chui, J. Bughin, J. Woetzel, P. Batra, R. Ko, and S. Sanghvi. Jobs lost, jobs gained: What the future of work will mean for jobs, skills, and wages. 2017.

[13] D. Ngo and Z. Bellahsene. Yam++: A multi-strategy based approach for ontology matching task. In *Knowledge Engineering and Knowledge Management: 18th International Conference, EKAW 2012, Galway City, Ireland, October 8-12, 2012. Proceedings 18*, pages 421–425. Springer, 2012.

[14] F. Osborne, A. Salatino, A. Birukou, and E. Motta. Automatic classification of springer nature proceedings with smart topic miner. In P. Groth, E. Simperl, A. Gray, M. Sabou, M. Krötzsch, F. Lecue, F. Flöck, and Y. Gil, editors, *The Semantic Web – ISWC 2016*, pages 383–399, Cham, 2016. Springer International Publishing.

[15] S. Pizard and D. Vallespir. Developing a taxonomy for software engineering education through an empirical approach. *CLEI Electronic Journal*, 23(2), 2020.

[16] V. Presutti, F. Draicchio, and A. Gangemi. Knowledge extraction based on discourse representation theory and linguistic frames. In A. ten Teije, J. Völker, S. Handschuh, H. Stuckenschmidt, M. d'Acquin, A. Nikolov, N. Aussenac-Gilles, and N. Hernandez, editors, *Knowledge Engineering and Knowledge Management*, pages 114–129, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

[17] A. A. Salatino, T. Thanapalasingam, A. Mannocci, A. Birukou, F. Osborne, and E. Motta. The Computer Science Ontology: A Comprehensive Automatically-Generated Taxonomy of Research Areas. *Data Intelligence*, 2(3):379–416, 07 2020.

[18] G. Stoilos, G. Stamou, and S. Kollias. A string metric for ontology alignment. In Y. Gil, E. Motta, V. R. Benjamins, and M. A. Musen, editors, *The Semantic Web – ISWC 2005*, pages 624–637, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg.

[19] C. Williams et al. Research methods. *Journal of Business & Economics Research (JBER)*, 5(3), 2007.