WIRTSCHAFTS
UNIVERSITÄT
WIEN VIENNA
UNIVERSITY OF
ECONOMICS
AND BUSINESS

Master Thesis

Identifying Key Concepts in Student-authored ontologies

Sofia Kovaleva

Date of Birth: 30.09.2000
Student ID: 12329182

Subject Area: Digital Economy
Supervisor: Prof. Dr. Marta Sabou

Date of Submission: 19.08.2025

Department of Information Systems & Operations Management, Vienna University of Economics and Business, Welthandelsplatz 1, 1020 Vienna, Austria

DEPARTMENT FÜR WIRTSCHAFTSINFORMATIK
UND OPERATIONS MANAGEMENT DEPARTMENT
OF INFORMATION SYSTEMS AND OPERATIONS
MANAGEMENT

Contents

List of Figures

List of Tables

Abstract

Ontologies as representations of concepts and their relations demonstrate a value for their possibilities of implementation, interoperability, and reasoning, Key concepts of ontology are a central or frequently used classes that reflects how domain knowledge is modeled. While the creation and reuse of ontologies have been extensively studied, less is known about the degree of conceptual overlap among ontologies authored independently by individuals with similar backgrounds.

The study investigates the phenomenon of overlap of concepts within ontologies engineered by students. It aims to investigate the phenomenon of key concepts overlaps, and whether such overlaps follow patterns of synonymity or cognitive structuring. The main research question explores the extent to which key concepts are reused across ontologies in two domains: music and movies.

To answer this, a semi-automated analysis pipeline was developed. It builds upon the key concept extraction method, developed by Silvio Peroni. The pipeline includes concept extraction, lemmatization, frequency analysis, clustering, and synonym detection. A total of 117 ontologies were examined through this analysis.

The findings reveal that students mostly use same concepts for description of different domains, most overlapping concepts tend to be short and have a certain level of abstraction. Across both domains, the most frequently encountered concepts were “genre” and “person”; more domain-specific overlaps were “album” in music and “actor” in movies. Analysis also demonstrated a stronger focus on human-related terms within the movie domain.

The study provides insights from both theoretical and practical perspectives of ontology engineering and holds potential implications for further research in concept alignment and the construction of core ontologies.

1 Introduction

1.1 Research Problem

Ontologies are described as representations of terms and concepts that are connected and refer to the same topic, forming a classification system [50]. They are used across various domains, such as philosophy, information technology, education and other. However, ontologies are most prominently employed in the field of the Semantic Web, where they serve as a foundation for capturing and constructing domain-specific language structures [6]. Conceptualization within ontologies serves as a cornerstone for enabling automated reasoning and the development of intelligent web services, which advances innovations to the next level [53]. Ontologies are also seen as a valuable player in shaping of possible "consensuses" - centralized structures that incorporate a wide range of concepts contributed by individuals. Those structures enhance interoperability and support the integration of knowledge both from and for diverse communities [34].

It is also essential to understand that the importance of ontologies lies not only in their existence, but in the underlying reason for their creation - namely, the human being. This importance is reflected in understanding how structures within ontologies align with human cognitive abilities, particularly in terms of how individuals conceptualize and construct such frameworks [7]. The underlying motives behind the cognitive processes involved in creating ontologies remain quite unexplored. This is reflected by a broader debate in neuroscience and philosophy, where it is difficult to determine what a "perfect" ontology would look like, given the differences in how people think and how individual brains process information [32]. This topic strongly depends not only on the relationships between concepts within ontologies, but also on the specific concepts that are used. It is these concepts specifically that can reveal how a person constructs an ontology [32].

The use of such patterns of concepts can be more clearly observed through the examination of key concepts (KC) within ontologies. The topic of understanding how KC are built is considerably important, as it may reveal critical connections that contribute to a deeper understanding of how the human brain operates when constructing taxonomies and semantic concepts. The question remains important not only for science, but also from an economic perspective. For instance, understanding the key concepts from which ontologies are constructed can lead to reduced time and effort required for their development [2].

Although the topic of key concepts in specific ontologies has been previously described [38], there are not many studies that compare ontologies created by individuals with similar backgrounds in a single domain. Despite existing efforts, the question of similarities in cognitive abilities during any creation process, as well as an ontology creation, remains unexplored [15].

This study aims to fill in the gap in existing knowledge on this topic and to examine potential connections between ontologies created by individuals of similar age and academic background. The study aims to address the problem through manual data analysis, Java-based program code, and a Python script. The problem will be examined using ontologies created by students of Semantic Web-related courses who developed ontologies within the same domains.

1.2 Research Questions

The question of how ontologies are created and the cognitive processes underlying their development remains as an acute problem. To examine this topic more closely, present study addresses one central research question alongside three sub-questions. By analyzing the semantic and lexical overlaps among independently developed ontologies, the study aims to determine whether a shared understanding of domain knowledge emerges despite individual variation in expression.

RQ: *To what extent do student-authored ontologies overlap in terms of the key concepts they use?*

In this study, it is important not only to examine the extent to which key concepts overlap within a single domain, but also to observe whether similar concepts appear across different domains. Addressing this question may offer insights into the degree of semantic proximity between ontologies on diverse topics and could contribute to answering whether it is possible to construct shared conceptialization ontologies that serve as comprehensive representations of domain knowledge.

SRQ1: *How does the degree of concept overlap vary across different ontology domains?*

Another question that can contribute to understanding which specific key concepts are shared relates to the observation that similar concepts in ontologies are often not complex terms or phrases, but rather simpler words that describe less specific notions - hypernyms [30].

SRQ2: *Are the overlapping key concepts predominantly general, high-level terms (hypernyms) with simple lexical forms (e.g., "person", "artist"), as opposed to more specific terms (e.g., "instrumental music")?*

The final sub-research question aims to understand whether key concepts are similar in meaning but differ syntactically - if they are synonyms of a sort. The use of consistent terminology in ontology construction is a critical issue, as it can support the future development of more generalized ontologies, facilitate ontology matching, and increase the speed and efficiency of knowledge integration processes [26].

SRQ3: *To what extent do overlapping key concepts share similar meanings, while expressed using different lexical forms (e.g., "musician" vs. "artist")?*

1.3 Methodology

To answer the research questions, the following methods, depicted in the Figure 1, are used: background research, data collection and analysis, application development and implementation, and empirical evaluation of the results. First, the background work and related literature review are explored. The methods used for further analysis in the study, as well as the tools applied, are justified. Then, the process of data collection is explained, including how and which data was collected, and which underlying patterns were identified during the collection and preliminary analysis. After that, the development of the application is described in detail, starting with an overview of existing work upon which the current solution is built. All added modules are explained, and a step-by-step overview of the process is provided. Finally, the results of the application are presented. They are explained according to the different types of analysis, which correspond to the research questions—focusing on key concepts themselves, their similarity (synonymity), and their overlap from a cluster perspective. The research questions are being answered through semi-automated analysis.

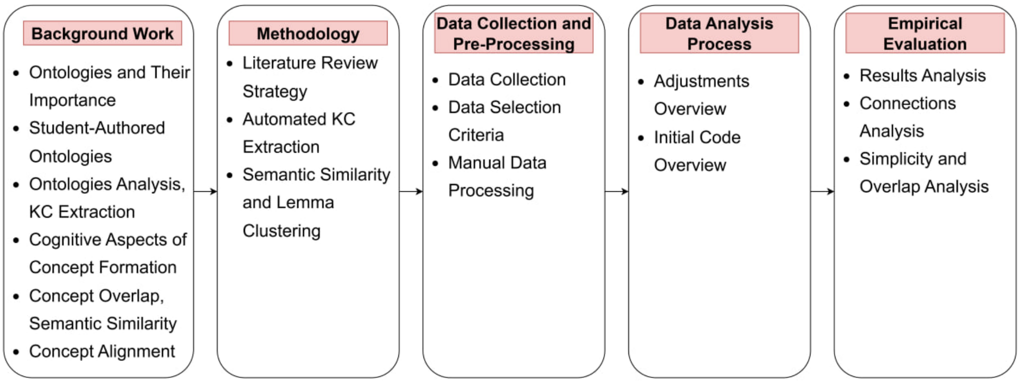
Background Work
Ontologies and Their Importance
Student-Authored Ontologies
Ontologies Analysis, KC Extraction
Cognitive Aspects of Concept Formation
Concept Overlap, Semantic Similarity
Concept Alignment
Methodology
Literature Review Strategy
Automated KC Extraction
Semantic Similarity and Lemma Clustering
Data Collection and Pre-Processing
Data Collection
Data Selection Criteria
Manual Data Processing
Data Analysis Process
Adjustments Overview
Initial Code Overview
Empirical Evaluation
Results Analysis
Connections Analysis
Simplicity and Overlap Analysis

Figure 1: Thesis Workflow Overview

1.4 Key Findings

Overall, a significant overlap in concepts can be observed within ontologies of specific domains — separately in music and in movies. The degree of overlap appears to be relatively equal across both domains, with a similar number of shared terms occurring within each.

The overlap across ontologies from different domains suggests that cross-domain arrangement is slightly possible. This means that, although limited in scale, it is feasible to create hybrid ontologies that spans across broader areas of knowledge and uses a shared set of terms.

Key findings also indicate that ontology engineers, in current study - students, tend to use short and easily understandable concepts, which supports Rosch's theory of basic-level categories [46]. Complex, multi-word expressions are rarely used and typically do not form synonymic pairs, whereas cognitively efficient terms appear frequently.

It was also found that a small number of synonyms were used during the creation of the ontologies. At the same time, limitations of the synonym matching method became evident, as the library applied (WordNet) does not account for domain-specific contexts.

1.5 Thesis Structure

The work is structured as follows:

In Chapter 2, the theoretical background is provided. An overview of ontologies, their role and importance is given. The decision to use student-authored ontologies for analysis is justified. This is followed by a description of methods for ontology evaluation and analysis that extend key concept extraction (KCE). Additionally, work discussess cognitive aspects of ontology engineering, which is complemented with a brief overview of concept alignment.

Chapter 3 describes the methodology of the study. It begins with an explanation of the literature review process, then explains the details of the applied automated process for key concept extraction, and concludes with an overview of the approach to semantic similarity and lemma clustering.

In Chapter 4, the justification and discussion for data collection and analysis are presented. The procedures for collecting of ontologies are described, followed by an explanation of the selected domains. Insights from manual data processing are then provided for both ontology groups.

Chapter 5 describes the data analysis stage. It begins with a theoretical overview supported by a sequence diagram for better understanding of the workflow. The chapter then presents the initial application code adapted from Peroni's work, followed by the adjustments and newly developed scripts created for the current study.

The empirical evaluation of the results is given in Chapter 6. Results from both domains are analyzed individually and jointly to address questions of conceptual overlap, similarity, and the theory of cognitive economy.

Chapter 7 provides an overview of the work and finalizes it. In it the discussion of the results, contributions of the study, limitations encountered, and potential directions for future research is given.

2 Background Work

This chapter discusses several important topics that are relevant to the current study. It begins in Section 2.1 by describing what ontologies are and why they are important for structuring and sharing knowledge in modern systems. Then, in Section 2.2 it explores how student-authored ontologies can be used to observe patterns in ontology construction and terminology use. The chapter also looks at existing methods for evaluating ontologies and identifying key concepts based on structure and usage in Section 2.3. Cognitive aspects behind the formation of concepts, including how people choose and categorize terms, are discussed in Section 2.4. Following this, an overview of how similar or overlapping concepts can appear across ontologies and how semantic similarity can be analyzed is given in Section 2.5. Finally, Section 2.6 explains why alignment of concepts is important for interoperability and ontology reuse, especially when dealing with independently created ontologies. Together, these topics form the theoretical background for the research approach, which is introduced in the next chapter.

2.1 Ontologies and Their Importance

Ontologies represent structured sets of data and the semantic relationships between them, situated within a specific domain [50]. In the current research, as well as in practice, ontologies are seen as a significant part of modern information systems. They are integrated within software infrastructures. They allow efficient access to domain-specific knowledge and facilitate repeated reuse across various applications. Such representation supports flexible and rapid interaction with complex data environments, as well as it allows knowledge distribution [11].

Ontologies are set within a specific type of structure. They consist of classes (concepts) - representations of domain entities; attributes, which resemble properties of classes; relationships, which represent the connections between classes; and rules (axioms), which determine classes and relationships behavior [51]. Their structures allow for wider knowledge usage and faster access to data.

Any ontology is developed according to its representational type. Domain ontologies, for example, primarily contain domain-specific terminology. Barry Smith referres to them as *"common-sense taxonomies"*, as they consist of concepts that are realistic and recognizable in everyday human experience, such as cat, human, or table [50]. In contrast, conceptual or top-level ontologies describe entities at a highly abstract and generalized level. These ontologies are often applicable across multiple domains due to their high-level nature and typically include fundamental concepts such as time, event, or object [50, 24]. Task ontologies are constructed to represent the knowledge needed to perform specific types of tasks, while application ontologies are used to implement such tasks in practice, often by integrating both domain and task ontologies [24]. Another caterogy is a recently emerged one, known as core ontologies, includes concepts that are, to some extent, contextualized but also general enough to be applied across various domains [14, 41]. They are positioned between top-level and domain ontologies, as is illustrated in Figure 2. This makes them well-suited for a wide range of use cases. Due to their intermediate level of abstraction and generality of concepts included, core ontologies promote interoperability and they support effective, low-cost, and error-resilient deployment across systems.

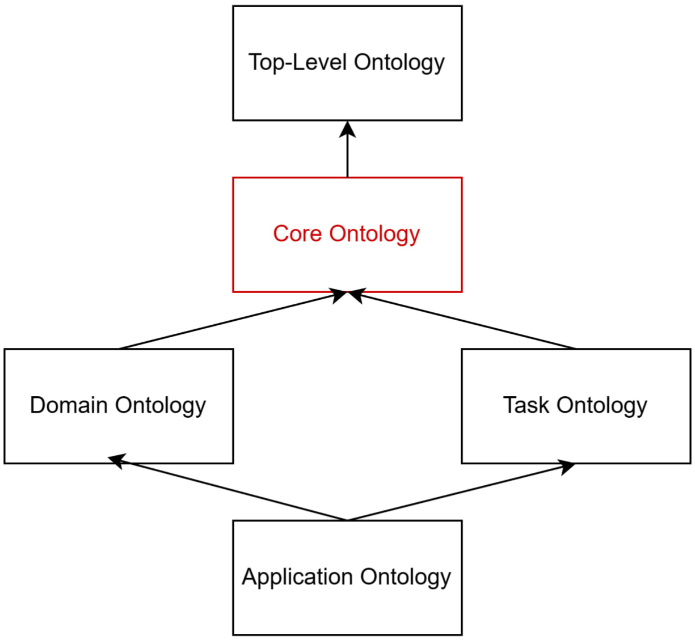

Figure 2: Types of Ontologies based on [16, 22]

Ontologies represent domain knowledge in a such formalized way, that supports more accurate and efficient use, interpretation, and exchange of information. Their importance is widely recognized across various domains. For instance, in the field of information technologies, they are most commonly utilized within the Semantic Web - a framework designed for the sharing and reuse of data [1]. Within the Semantic Web, ontologies define connections between terminologies and serve as a source of information. They are most commonly presented using a variation of the Resource Description Framework (RDF) as triples in the form of *subject-predicate-object* [13, 57], which can be further formalized using the Web Ontology Language (OWL) to define instances of concepts, relationships, and logical rules [33]. These concepts are widely known and used as they have been developed and standardized by the World Wide Web Consortium (W3C) [57].

The Semantic Web enables the usage of ontologies in a variety of real-world contexts. One of the most outstanding use-cases is in enhanced search technologies. These technologies go beyond traditional keyword-based search. In contrast, they utilize ontologies to understand the meaning and context of queries and data [18]. This does not only support percise use of queries, but also accelerates access to context-aware answers, which is an essential advantage in today's fast-paced digital environment. The application of ontologies in this context is also evident in domains such as healthcare. One of such examples is AstraZeneca's involvement in the LarKC project. The project focused on early hypothesis testing by employing ontology-based search via Ontotext to enhance the relevance and contextual accuracy of retrieved results [35, 4].

This examples supports the potential of ontology-based search for automated reasoning. Systems with automated reasoning can conclude new knowledge based on existing data and ontological relationships. This reasoning supports more advanced machine functionality, including memory-based retrieval, semantic understanding, and continuous logical processing within specific domains [39]. All of this is made possible through ontology-based classification and structured knowledge representation.

Both enhanced search and automated reasoning can benefit from semantic interoperability, which serves as a foundation for effective data transmission across diverse systems [28]. Ontologies within the Semantic Web are considered as a basis for promoting interoperability, as they provide a structured vocabulary that supports consistent understanding of information. For example, a semantic-based approach, applied in large-scale software systems, such as Internet of Things (IoT), has demonstrated a positive impact on device communication [43]. Interoperability can be further achieved through the reuse and scalability of ontologies.

These considerations point to a single potential solution: the development of a large-scale core ("consensus") ontology, or a set of such ontologies, that is robust, flexible, and applicable across multiple contexts. Such ontologies would be composed mostly of key contextual concepts, as such concepts make them both efficient and reusable.

To assess the feasibility of this approach, the present study investigates whether it is possible to systematically identify the core contextual terms embedded in ontologies created by individuals.

2.2 Student-Authorized Ontologies as a Research Focus

Student-authored ontologies are an ideal choice for the current study, as they provide an extensive view and amount of data from people with similar backgrounds in ontology creation, but differing levels of domain knowledge. Student-based ontologies can offer valuable information about the thinking process of students who understand ontology engineering and actively engage in building ontologies, making them a strong example for observing word use and terminology selection.

Since the ontologies they created are not related to their specific fields of study, we are able to observe an informal, unconstrained and intuitive pattern of ontology construction directly from the source. This also allows us to identify whether any consistent patterns in terminology usage emerge, which could support the development of core ontologies in specific domains.

The use of such ontologies may also contribute to understanding how cognitive patterns function during ontology construction, as well as how individuals conceptualize and manage knowledge.

2.3 Ontologies Evaluation, Analysis and Key Concepts Extraction

To understand the topic of key concepts, their role and their contribution to the development of new ontologies, it is first important to examine how ontologies are evaluated and how key concepts are typically identified.

The evaluation of ontologies can significantly contribute to understanding their overall value and quality, as well as the relevance of the concepts used within them. Ontology evaluation is conducted differently depending on the type of ontology and the specific purpose behind the evaluation. Those methods can include looking at criteria such as consistency, completeness, and conciseness, as well as identifying possible taxonomic errors like circularity, redundancy, and incompleteness [19].

Most evaluation approaches typically focus on the connections between terms, as well as the consistency and terminology of entities. The OntoClean approach, proposed by Guarino and Welty, examines the taxonomic structure of ontologies and emphasizes that rigidity, identity, and unity are crucial properties for evaluating ontological terms [25]. Other methodology called OOPS! focuses on error detection in ontologies. In their work, Poveda-Villalón et al. propose that "pitfalls", or mistakes made by ontology engineers, can be classified into structural, functional, and usability-profiling dimensions, each of which can have an importance level [40]. A more important approach for the current study, a metric-based approach OntoQA, has been made for evaluation of ontologies for their reuse, looks into schemas of ontologies, knowledge base metrics and class-level metrics [52]. Through its metrics, OntoQA highlights the most populated or linked classes, which helps in finding key concepts and understanding their value. While evaluation techniques ensure that ontology is well-structured and consistent, analytical approaches allow for deeper insight into how concepts are used, related, and prioritized.

A graph-based ranking algorithm LexRank is a document text analysis method that examines graphs and their centrality. Though the analysis is applied to texts rather than ontologies, this method, just like ontology descriptions, relies on graph structures, and thus provides insight into how centrality can be discovered within a semantic framework. The method allows for the extraction of important units based on their structural position in the graph [17]. The Concept Appearance Ranking (CARRank) method, in turn, approaches semantic analysis from a slightly different perspective by estimating the importance of concepts and relations within an ontology. Proposed by Wu et al., the method uses an importance ranking model that analyzes the graph representation of an ontology. Although the type of analysis differs, it the general approach of using graph representations is a commonality with LexRank analysis. Though semi-user-dependent, CARRank helps identify the most valuable concepts and relationships within an ontology, which can then be reused [58]. Another analysis type, A Dual Walk based Ranking model (DWRank), focuses primarily on the relationships between concepts to identify the most relevant ones. Similar to LexRank and CARRank, it uses graph-based approach. It introduces a hub score and an authority score to measure intra- and inter-ontology connections and potential for reuse. Combined with text relevance, these scores are used to rank ontology concepts. This supports the semantic reuse of ontologies and improves their discoverability [8].

All of these methods can assist in describing well-established ontologies as well as in identifying potential defects. Ontology evaluation can not only reveal error-prone approaches in ontology engineering, but also help detect weak concepts and connections, which are essential factors for the development of a high-quality ontology, such as a core ontology. While traditional evaluation focuses on structural and logical quality, and described analysis methods can highlight the most significant concepts within an ontology, the identification of key concepts can further support this process.

While many studies focus on techniques for extracting key concepts from texts or academic papers to support ontology construction, it is also important to recognize that extracting key concepts from existing ontologies within a specific domain can significantly contribute to ontology engineering. The key concepts of ontologies describe their most fundamental and meaningful elements. The extraction of key concepts from various ontologies within the same domain can reveal how their authors approach conceptualization and may help identify specific classes that can later be used to construct a precise core ontology.

One of the foundational papers, which is a direct inspiration for the current study, is Peroni's work on key concepts identification. In contrast to other evaluation and analysis methods, this approach is fully automated. Peroni's algorithm identifies key concepts by examining several aspects, including the local and global density of a term (based on the depth of neighboring nodes), as well as coverage, which ensures that the selected concepts provide a broad representation of the entire ontology. The approach builds on the theory of natural categories, which are more commonly used due to their simplicity and concreteness in description [46]. In a later revision of the work, a popularity criterion was also introduced to account for widely known concepts that might be overlooked by lexical simplicity or density-based selection [38]. As an outcome, the algorithm provides a list of key concepts identified within an ontology, which in turn supports knowledge distrubition and the reuse of ontologies.

During the literature research few if any studies that focus on the extraction of key concepts from existing ontologies were found. This builds a gap in the research leading to inability to production of proper domain, core and upper-level ontologies. This topic can highly contribute to the building of core ontologies and therefore support faster and more available knowledge usage.

2.4 Cognitive Aspects of Concept Formation

Ontology entities generally consist of terms that have been categorized. Therefore, key concepts can be seen as the result of categorization and conceptualization of those terms. To further understand how key concepts as terms are formed in the human mind, it is important to address the psychological and lexical aspects of their formation.

In her work [46], Eleanor Rosch states that human conceptualization is guided by two principles: cognitive economy and perceived world structure. The first principle suggests that the mind typically aims to minimize effort while maximizing informational gain. This means that people tend to prefer terms that cover a broad area of knowledge with the least cognitive effort.

The second principle states that conceptualization depends on the structure of the perceived world and therefore tends to reoccur naturally. When forming concepts, the mind identifies correlations between terms and constructs familiar and recognizable concepts that relate to the real, known world and cultural context. Rosch implies that people usually rely on categories - simple terms with a basic level of abstraction. The most effective of these categories possess distinct "cues": mental associations that allow individuals to recognize an object and distinguish it from others. Based on this theory, word "chair" represents basic-level category that is most applicable from a cognitive perspective for ontology building, rather than "furniture", which is too vague or cumbersome, and "kitchen chair", which is too specific.

The work of Hampton [27] explores Rosches theory further and suggests that concepts are prototypes with four cognitive features: vagueness, typicality, genericity, and opacity. Hampton explains that people often struggle to precisely define the categories they use, as their intended meaning does not always align with used definitions. He also suggests that some categories are used more frequently and therefore percieved as more typical than others, and that generic descriptions, though common, do not necessarily apply to all members of a category. It is also stated that people are prone to error when engineering categories, as they tend to use intuition rather than logic and rule application, which leads to opacity. Those are one of the most vital concepts for ontology engineering, as they allow to understand how brain activity is involved in the development process.

Apart from the precise definition of categories, people also often struggle with finding the right phrasing for them. In such cases, synonyms, hyponyms, or hypernyms may be used to describe the same concept from different perspectives. Synonyms are words that convey similar meanings, but differ in lexical form. Hyponyms and hypernyms are opposites: a hyponym refers to a more specific term, while a hypernym denotes a broader or more general category. For example, “auto” is a hyponym of its hypernym “vehicle,” while “car” may function as a synonym of “auto".

While Rosch’s study shows that people tend to process simpler and more intuitive terms, Chaffin and Glass support this idea through experimental findings. In their work [10], participants were asked to evaluate the truth of simple categorical statements involving either hyponym or synonym pairs. The results showed that people tend to comprehend hyponyms faster and with greater ease, as they are cognitively simpler and more directly structured. This supports the idea that simpler and more specific terms are naturally preferred, which helps explain patterns in word usage during ontology engineering.

A more practical perspective on lexical variation in ontologies is offered by

Kwak and Yong [30], who show that relying solely on exact lexical matches can overlook other possible concept alignments. Researchers developed a novice ontology matching method called Super Word Set Similarity (SWS). Within the SWS they incorporated score system for semantic relations such as synonyms, hypernyms, and hyponyms. The results showed how similar meanings may be expressed through different terms, outperforming tools more commonly used, such as COMA++ and LOM, which both process on WordNet. Their findings empathize the importance of lexical variation recognition when evaluating concept overlap, particularly in relation to whether concept similarity is expressed through alternative but semantically related terms, as addressed in SRQ3.

Ontology engineering is directly connected to the development of lexical terms and the words people use in everyday life. The question of employing cognitively acceptable, natural, and human-friendly terms in ontologies is actively discussed within the ontology engineering community. It is true that both the usability and the potential for knowledge dissemination increase when ontologies are built using simpler and more familiar and accessible terminology.

The importance of guiding ontology engineering in a cognitively simple and user-friendly manner is thoroughly discussed in the work of Daniel Schober et al. [49]. The study recommends constructing ontologies using broadly meaningful terms — entities that can be interpreted across different contexts without losing clarity. At the same time, concept names should follow specific and consistent formatting rules, such as, for instance, being written in singular form. To maintain simplicity, acronyms and abbreviations should remain abbreviated, rather than expanded. Additionally, it is suggested to avoid negative prefixes such as “non-” or “un-”, as they may be misunderstood and block interoperability. Overall, the study emphasizes on a human-centered approach to ontology development, which helps reduce errors, improves processing efficiency, and facilitates reuse across domains. This supports investigation of SRQ2, which explores possibilities of concepts being more simplified and high-level, as such are easier to process.

The use of naturally occurring terms is explored in the work of Jung An et al. [3]. In their study, the naturalness of a term is described as a measurable metric based on three key criteria: its frequency of usage, lexical simplicity, and its availability in widely used sources, such as Google or general dictionaries. Lexical simplicity in their work is measured by the number of words in a concept name. The authors analyzed several well-known ontologies to evaluate these factors and conducted statistical tests, including t-tests and ANOVA, to compare naturalness scores across different datasets. Their findings suggest that higher naturalness improves the overall quality and interpretability of ontologies. They conclude that replacing rare or unfamiliar terms with more frequent and recognizable synonyms can enhance usability. Overall, their work shows that naturalness of terms leads to better comprehension, usability, and knowledge sharing in ontology engineering.

It is well-known that people aim to use efficient and often simpler, more general terms in both speech and decision making. During conceptualization, people frequently use hypernyms, hyponyms, or synonyms to express the meaning. This becomes even more evident in cases where exact terminology is difficult to define. Additionally, people tend to prefer naturally occurring, frequently used terms that feel familiar in a common language setting. These factors significantly influence how ontology concepts are formed and named, and this understanding supports the relevance of the current study in analyzing student-authored ontologies and assessing whether the theoretical patterns identified by researchers are reflected in practice.

2.5 Concept Overlap and Semantic Similarity

As discussed previously, semantic interoperability is a crucial aspect of ontology engineering. Analysis of conceptual thinking shows that people tend to use semantically similar terms interchangeably, which often leads to concept overlap. This, in turn, can introduce errors into ontology structure and slow down knowledge distribution. Therefore, it becomes essential to understand how to address the use of semantically overlapping terms and minimize their impact.

One approach to addressing this issue is presented in the work of Santos et al. [48]. The authors discuss the challenge of overlapping concepts or classes in ontologies. They note that such overlaps can reduce semantic clarity, obstruct interoperability, and complicate ontology reuse and knowledge integration. To mitigate these effects, they propose a method for identification of structurally and semantically important concepts, which represent core knowledge within ontologies. The method involves three phases: decomposition, identification, and quantification of overlap. Their work introduces specific metrics to assess overlap, such as label similarity, shared properties, and common hierarchical structures. Unfortunatelly, paper discussess diverse areas of overlap with different concepts concentration. In some domains concepts are densely packed, while in others they are of low density, which makes study imbalanced. While the current work does not explicitly focus on imbalanced domains, the insights offered by this method remains applicable, as student-authored ontologies vary in coverage and conceptual density depending on the author's background and focus.

Concept overlap is a topic that is closely connected with the semantic similarity of concepts. Similar to overlap identification, there are several ways to detect semantically related concepts. In the work of Chandrasekaran and Mago [12], four major types of methods are described: knowledge-based, corpus-based, deep neural network–based, and hybrid approaches. Knowledge-based methods rely on hierarchical structures, such as WordNet, to compare terms based on their taxonomic distance or shared attributes. Corpus-based methods extract semantic meaning from patterns in large text corpora using statistical techniques like co-occurrence and topic modeling. Deep neural network methods, though highly effective, depend on complex language models, such as BERT, and are difficult to interpret or apply manually. The current work can benefit from the study of Chandrasekaran and Mago by partially implementing knowledge-based and corpus-based principles, such as examining label paths and comparing definitions to see how close the meanings are.

This approach helps to uncover concept alignments even when terms differ in form but stay close in meaning. The authors highlight how semantic similarity can help to overcome lexical variation, especially when comparing ontologies built separately. In the context of student-authored ontologies, applying these principles may reveal understanding behind concept creation or if terms are used differently. Proposed work discovers the topic of similarity via usage of WordNet, which was mentioned above.

Most of the works that address concept overlap have not implemented any automated or semi-automated analysis and were mostly based on manual examination or focused on the cognitive side of ontology and concept formation. For example, the work of Wu et al. [58] focused on identifying important concepts and relations within ontologies, but did not provide a framework for systematic lemmatization or clustering. Similarly, Taye's overview [53] offers foundational theoretical insights into the Semantic Web and ontology usage, but lacks empirical or computational analysis of overlapping concepts across domains.

2.6 Importance of Concept Alignment

One of the key goals of ontology engineering is to not only model domain knowledge accurately, but also to ensure that ontologies can be integrated across various different contexts and systems. This becomes especially relevant when ontologies are independently developed, or come from diverse domains, as in the case of student-authored ontologies. Semantic interoperability, enhanced knowledge reuse, and improvement of ontology matching and integration are advantages of concepts alignment. When concepts are aligned, systems can recognize them as referring to the same or similar ideas, which reduces redundancy, conflict, and fragmentation of knowledge, as well as it supports ontologies reusage.

Alignment of concepts is closely connected to the ontology alignment. Granitzer et al. [21] explain that alignment is a challenging task and it can be obstructed, especially when people use different ways to describe the same idea. Even if two concepts have the same meaning, they might appear different due to variations in the usage of vocabulary or their structure. This makes it difficult for systems to compare and understand them correctly. The authors show that fully automatic alignment tools often produce errors and cannot be fully relied upon. Authors state that alignment cannot be fully automated, but with manual processing it can lead to positive, especially when comparing concepts that have same meaning, but differ in syntax.

This idea is explored further in the work of Ardjani et al. [5], where ontology alignment is defined as the process of finding parallels between entities from different ontologies, such as classes, properties, or instances. Authors describe several ways to approach alignment. They include label similarity, hierarchical structure comparison, instance overlap, and use of external knowledge sources. Authors state that alignment methods work best when used together, as different strategies can cover different aspects of semantic closeness. The work states that ontology alignment is one of the most important steps to achieve interoperability, as it allows systems to work with independently built ontologies. This is a relevant remark in the case of student-authored ontologies, where same ideas might be described differently, but can still be aligned through shared meaning.

Ontology reuse is widely supported by concept alignment and brings benefits to the ontology engineering community. According to the work of Carriero et al. [9], ontology reuse can be performed directly or indirectly, depending on the field of use, as well as the purpose and structure of the reused ontology. Direct reuse implies importing classes or modules from existing ontologies, while indirect reuse is focused on adopting conceptual structures or patterns without direct import of terms. Ontology reuse strategies may vary based on how the developer intends to reuse existing content: through standardisation, popularity, or cognitive/conceptual alignment. When choosing a reuse approach, it is important to consider four key dimensions that influence ontology reuse practices overall - ontology selection, access, integration, and reuse implementation. Authors also note that alignment of concepts helps to make ontology reuse more effective, especially when ontologies are being integrated or merged. At the same time, reuse strategies, such as indirect reuse, depend on the ability to find similar or close in meaning concepts, which shows that alignment and reuse are strongly connected and support each other.

2.7 Summary

The importance of ontologies and their role in structuring domain knowledge, discussed in Section 2.1, points out the relevance of understanding how concepts are built and named. Section 2.2, where terminology use tends to follow informal and intuitive patterns, shows the significance of concepts engineering in student-authored ontologies. The need to evaluate these ontologies and extract their key concepts is addressed in Section 2.3, which outlines both traditional evaluation methods and graph-based approaches for identifying concept importance. Section 2.4 further explains how cognitive processes influence the formation of concepts, supporting the idea that users naturally prefer simple, familiar, and general terms. The issue of concept overlap and semantic similarity, explored in Section 2.5, connects directly to the challenges of lexical variation in ontology construction. Finally, the relevance of aligning concepts across independently created ontologies, as discussed in Section 2.6, forms the foundation for the methodological approach developed in the following chapter.

3 Methodology

This chapter explores and justifies the usage of different technologies applied throughout the study. First, in Section 3.1, it introduces the method of literature review used to explore the background of the current work. Then, the technologies applied during the development phase of the study are described in Sections 3.2 and 3.3. In Section 3.2, the enhancement of an application for key concept extraction, lemmatization, and analysis is explained. Further analysis of lemmas through clustering and visualization, used to identify similarities across ontologies, is discussed in Section 3.3.

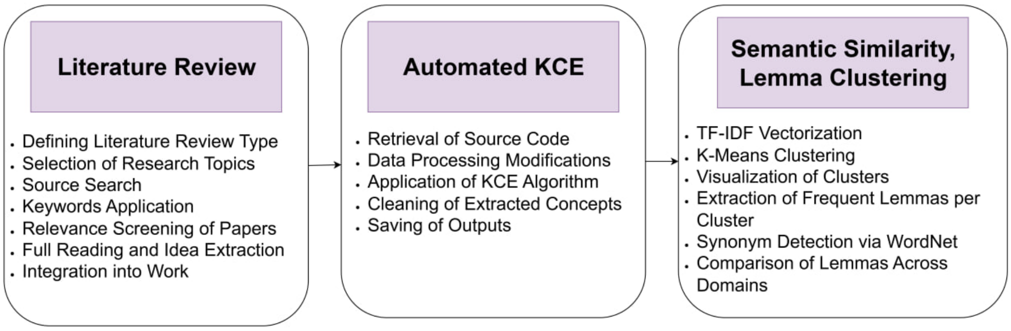
Literature Review
- Defining Literature Review Type
- Selection of Research Topics
- Source Search
- Keywords Application
- Relevance Screening of Papers
- Full Reading and Idea Extraction
- Integration into Work
Automated KCE
- Retrieval of Source Code
- Data Processing Modifications
- Application of KCE Algorithm
- Cleaning of Extracted Concepts
- Saving of Outputs
Semantic Similarity, Lemma Clustering
- TF-IDF Vectorization
- K-Means Clustering
- Visualization of Clusters
- Extraction of Frequent Lemmas per Cluster
- Synonym Detection via WordNet
- Comparison of Lemmas Across Domains

Figure 3: Methodology Workflow

3.1 Literature Review Strategy

Current work incorporates literature review within the discussion of previous works and background to the topic. For current study, an integrative literature review was conducted. Integrative review proposes a field for further research studies, as it encorporates analysis of both theoretical and practical papers [56]. In this study, theoretical findings were supported with empirical research outcomes. These were then compared across different contexts, and their combined insights were used to guide the focus of the work. This helped to provide an overview of the field, obtain current status of researches in similar topics, and offered orienteers for some tools selection as well as further development.

For finding background information, a scope of papers was chosen. Most of the papers, which included research, surveys, or other practical implementations, were primarily taken from the last 20-25 years, as this period marks the rise of the Semantic Web. A significant number of papers on concept and ontology alig nment were selected from the last 5-10 years, which shows that the topic is still evolving and remains relevant. Papers regarding the theoretical background of ontologies, their use and importance, as well as cognitive aspects behind them, date back to the 1960s and later, which shows that the topic has been discussed for many decades.

To find background literature, sources were mainly accessed via Google Scholar, ResearchGate, and the WU Library [20, 44, 55]. Search terms such as “ontology”, “concepts”, “alignment”, and “cognitive” were used. Since the background spans several topics, more specific keywords were applied to each area. This included research on "core ontologies", "ontological interoperability and reuse", "categorization and conceptualization", and tools such as "WordNet" or clustarisation ones via "K-Means", "DBSCAN", and "Hierarchical Clustering", as well as cluster validation techniques like the "Elbow Method" and "Silhouette Score". Little to no non-academic websites were used, as scholarly literature was broadly available. However, websites such as the Semantic Web portal served as a main source for information on that topic. Technical websites related to Python and Java were also referred to, as they were relevant for specific functions and tools used in the implementation.

First, for each background topic, a scope of around ten papers was identified. Then, their abstracts were read through. Criteria of exclusion were not strictly defined, as the area of knowledge is widely discussed and covers multiple perspectives. For papers regarding practical implementations, those conducted in the last 5-10 years were more likely to be included. For theoretical information, book chapters or more frequently cited papers were preferred over others. If the information in the abstract corresponded to the topic, the paper was read in full, and key ideas were collected and incorporated into the current work.

3.2 Automated Key Concepts Extraction

Automated key concept extraction was based on the existing work of Peroni [37]; therefore, the use of Java as the implementation language was not a deliberate methodological choice, but rather determined by the design of the original tool. In general, Java is widely recognized for its strong performance in ontology-related tasks due to its compatibility with ontology libraries such as OWL API [29], its robustness in managing large data structures, as well as the portability across systems. Several studies highlight Java’s advantages in semantic web development and ontology engineering [51, 58].

The source code of the tool, which is openly available on GitHub [37], was downloaded and adapted for analysis conducted in the current study. During the enhancement process, additional technologies were integrated into the pipeline. In particular, Stanford CoreNLP was used to normalize extracted key concepts by converting them into their lemma forms. Stanford

CoreNLP is a powerful tool that operates through so-called `Annotators` and offers various functionalities for the user. For instance, it can be used as a tokenizer, sentence splitter, or lemmatizer, as it was applied in the case of the current study. Standford CoreNLP is easy to use, as it is implemented as a Java API, compared to heavier frameworks, such as UIMA or GATE [31]. The lemmatization step performed by CoreNLP was essential for the later clustering process, as it helpes to ensure consistency across key concepts and reduce variation.

3.3 Semantic Similarity and Lemma Clustering

The third part of the study takes a step away from the Java programming language and uses Python instead for deeper analytical processing of the Java-based outputs. Python is one of the languages that was extensively used by the author during both bachelor and current master studies, which contributed to the choice of programming language for analysis. Another reason for choosing Python as the primary language for analysis is that it provides access to a wide range of libraries, such as `scikit-learn` [36]. This library offers tools for clustering, which is essential in the current work.

From this library, the `TfidfVectorizer` method was used. It enabled the calculation of term frequency and inverse document frequency (TF-IDF) scores, which are commonly used to determine the importance of a term in a specific document relative to a larger corpus. TF-IDF combines two components: term frequency (TF), which measures how often a word appears in a document, and inverse document frequency (IDF), which assigns lower weights to terms that occur more frequently across multiple documents [42].

Clustering is a method that groups similar items based on shared features [45]. In text analysis, clustering allows finding of related concepts across different documents. In the current work it works across multiple student-generated ontologies by grouping them based on TF-IDF representations. This is needed to confirm or challenge our assumption that ontology creators tend to use semantically adjacent concepts when building ontologies. It also helps uncover potential interconnections between sub-themes within those ontologies.

For this study, the k-means algorithm was used, as this is a well-established clustering technique that splits data into k groups, with each group represented by the mean of its elements [45]. K-means is a highly efficient method that effectively works with tasks of clustering [59]. To ensure reliable cluster selection, it was decided to evaluate cluster quality using both the Elbow method and Silhouette scores. The Elbow method is based on the Sum of Squared Errors (SSE) and identifies the point at which, with a growing number of clusters, their usability decreases. This is called a “knee point”, which is visually seen as a break point on a plot, and is needed for a balance between amount of clusters and data within them [54]. The Silhouette method, introduced by Rousseeuw [47], provides a measure of how well each object fits within its assigned cluster compared to other clusters. It reflects both cohesion (within-cluster similarity) and separation (between-cluster difference). As an outcome, it gives a value between -1 and +1, which shows the quality of cluster assignment for each data point.

The raw data used in the current analysis are provided in the form of an Excel file. Therefore, libraries such as `pandas` and `NumPy` were used, as they offer tools for data loading and manipulation, as well as other numerical operations. For the visualisation of clustering results, `matplotlib` was applied to generate the necessary plots.

4 Data Collection and Pre-Processing

This chapter presents the work on data collection and pre-processing, which constitutes the basis for investigating the research questions of this thesis. In Section 4.1, the process of collecting ontologies on two main topics, music and movies, is described and justified. In the next Section 4.2 the criteria for selecting ontologies on these topics are explained, including the terminology used for searching and the exact number of files collected. Following that, the description of manual data processing is presented in Section 4.3. The procedure is described separately for music and movie ontologies in two subsections, with details on the problems encountered and initial patterns observed in student ontology engineering, before any automated analysis was conducted.

4.1 Data Collection

Data collection process constitutes as one of the most important and fundamental steps of the research. In order to explore the possibility of collective thinking, it was decided to consider for study those ontologies, which were developed on the basis of widely popular and broad topics.

The decision to focus on broad, decentralized subjects was made to ensure flexibility in the results and to potentially identify integrations between different aspects of ontologies created by various individuals. Specifically, the following topics were selected:

- Music;
- Movies.

These topics were chosen as they can share thematic similarities and structural patterns between themselves.

The ontologies were collected using a source provided by the Wirtschaftsuniversität Wien (WU) through the "Ontology Search" web platform [1]. This service includes ontologies, which were collected by the Semantic Systems team at the university. The ontologies were developed by students of either WU or Technische Universität Wien (TU) during their academic studies.

All ontologies were anonymized before being uploaded to the platform; thus, it is not possible to determine which student or academic program contributed to each ontology. Each ontology on the platform includes three attributes: name, description and file. While name and description are self-explanatory terms, the file refers to the ontology document, provided in either the OWL or TTL formats, which are the two most commonly used formats for ontologies.

4.2 Data Selection Criteria

As it has been previously outlined, ontologies for two topics were collected. These topics were selected based on several considerations.

First, it was important to shift the focus of the study towards a wide range of terms that would neither restrict the amount of collectible data nor introduce overly specific descriptive terminology. Instead, the focus was placed on broader concepts within each domain, as these tend to offer greater value for subsequent analysis [23]. In the selection process, various topics were considered, all of which were deliberately broad, including: "food", "books", "university", "tourism", "travel", "music", "movies", "films", "art", and "theater". Some of these topics could have included potentially related concepts; however, the number of ontologies found for them was too low to support a meaningful analysis.

Two topics,"music" and "movies", yielded the highest number of available ontologies. During the search process, topics such as "theater", "films", and "art" were explored. However, for some terms, such as "theater" and "art", no related ontologies were found. Interestingly, for the term "films," significantly fewer ontologies were retrieved compared to the term "movies" — 11 and 72 ontologies respectively. Some of the ontologies found under the "films" query were also included among those retrieved under the "movies" search.

Second, these topics were selected due to their potential to exhibit similarities between them. It is essential for the study to examine not only commonalities in patterns of thinking within a single topic, but also to investigate potential correlations across distinct topics. Music can be connected to specific movies, and both domains may share complementary genres, such as "fantasy" or "horror." Furthermore, both topics exhibit diversity and may include references to individuals, who contribute to both fields, such as Elvis Presley or Jared Leto.

For the purposes of this study, all accessible ontologies related to the selected topics were included. In total, 60 ontologies were gathered for the "music" query, and 72 ontologies were collected for the "movies" query.

4.3 Manual Data Processing

Manual data processing included the downloading and collection of ontologies, as well as the documentation of their properties (name and description) and the subsequent analysis of the concepts contained within these ontologies.

For the analysis of concepts, a key concept code was initially applied, as described in Chapter 5. This approach allowed the extraction and storage of key concepts for each ontology into dedicated text (TXT) files, which further enabled a detailed manual review of the collected data.

The further approach to data exploration remains semi-manual, as modifications within ontologies or decisions regarding their exclusion were made following a review of the data produced by automated operations: whether concept extraction, lemmatization, or subsequent analysis.

4.3.1 Music Ontologies

From the "Ontology Search" platform [1], a total of 60 ontologies related to the topic "Music" were retrieved. The name and description of each ontology were recorded in an Excel file. Within this file, each ontology was assigned an ordinal identifier composed of the prefix "music" followed by a sequential number (e.g., "music1"). The use of the spreadsheet facilitated the tracking of items throughout both the manual and automated analysis stages. During this process, each ontology was also downloaded. However, eight ontologies - namely "music5", "music6", "music21", "music41" and "music43" - could not be downloaded due to file inaccessibility on the platform. Therefore, the total number of available ontologies was reduced to 55.

Following the transformation of the ontologies, key concepts were extracted using Peroni's code [37], with modifications made in this work. These modifications allowed saving the outputs into individual files, each corresponding to the respective ontology. After this each file was individually examined. During this examination, inconsistencies in the key concept outputs were identified. For example, in the case of ontology "music60" an error was observed, where each key concept contained the redundant prefix "musical". To ensure data consistency, a manual correction was performed by removing the prefix from each key concept entry. Additionally, ontologies "movies69" and "movies71" were found to be duplicates, both containing unreadable or hashed data, such as "rdvlmmm7xmbejtlwoeurfvp". Consequently, these ontologies were excluded from further analysis. Unfortunately, the issue of ontology duplication was encountered more than once, as will be discussed in greater detail in Chapter 5. Therefore, for the purposes of the automated analysis, only 52 ontologies, which were percieved as unique, were considered.

4.3.2 Movies Ontologies

From the "Ontology Search" platform [1], a total of 72 ontologies related to the topic "Movies" were retrieved, which is more than for ontologies on the "Music" topic. The name and description of each ontology were recorded in an Excel file, same approach as with "Music" ontologies. Within this file, each ontology was assigned an ordinal identifier composed of the prefix "movies" followed by a sequential number (e.g., "movies1").

During this process, each ontology was also downloaded. However, five ontologies — "movies2", "movies16", "movies29", "movies30", "movies60", "movies64", "movies69" and "movies71". Those could not be downloaded due to file inaccessibility on the platform. Consequently, the total number of available ontologies was reduced to 64.

Following the transformation of the ontologies into key concept files, these files were thoroughly examined. During this examination, inconsistencies in the key concept outputs were identified. For example, in the case of ontology "movies57" an error was observed, where each key concept contained the redundant prefix "musicsontology". This led to the assumption that the ontology may be related to music rather than movies, despite being found within the scope of the "movies" category and having a name and description that support this. To ensure data consistency, the ontology was excluded from the current analysis. Similar to music ontologies, in movies ontology "movies54" an error of unreadable or hashed data, such as "rda24784ws9o2frwukmuzqv", was found. Consequently, this ontology was excluded from further analysis. Therefore, for the purposes of the automated analysis, only 62 ontologies, which were percieved as unique, were considered.

4.4 Summary

Overall, 55 music and 62 movie ontologies were considered for use in the current work. Manual analysis was conducted for all of them at different stages, beginning with the extraction of key concepts. After verification, some ontologies were excluded from the study, as they either lacked substantial information or were duplicates.

5 Data Analysis Process and Implementation

The current chapter explores the development of the process for ontology analysis from start to finish. The process consists of sequential steps: first, ontologies are collected and manually reviewed based on their descriptions and names. They then undergo key concept extraction, lemmatization, and subsequent analysis, including clustering. After each automated step, a semi-manual data review is conducted to exclude duplicates or ontologies with processing failures from further consideration. Finally, the results of the analysis are examined.

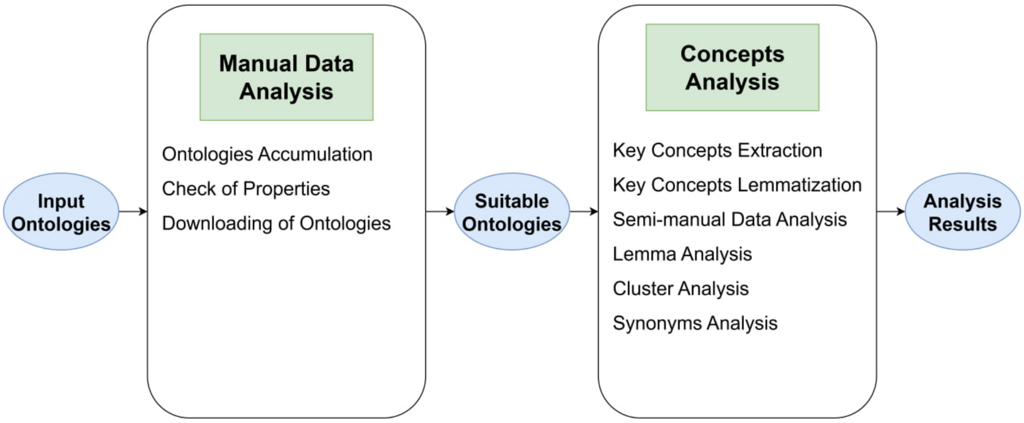

Figure 4: Workflow of Data Analysis

The algorithm's operation is illustrated in Figure 5. The process proceeds as follows: uploaded ontologies undergo key concept extraction, after which the extracted concepts are saved into individual files. From these files, the concepts are transformed into lemmas, which are then stored. Subsequently, analysis is performed using the lemmas by identifying those that appear in more than 50% of the ontologies and conducting a cluster analysis. For convenience, before clustering analysis the data is first exported to an Excel file. Following the Figure 5, algorithm is explained in detail in Chapter 5.1.

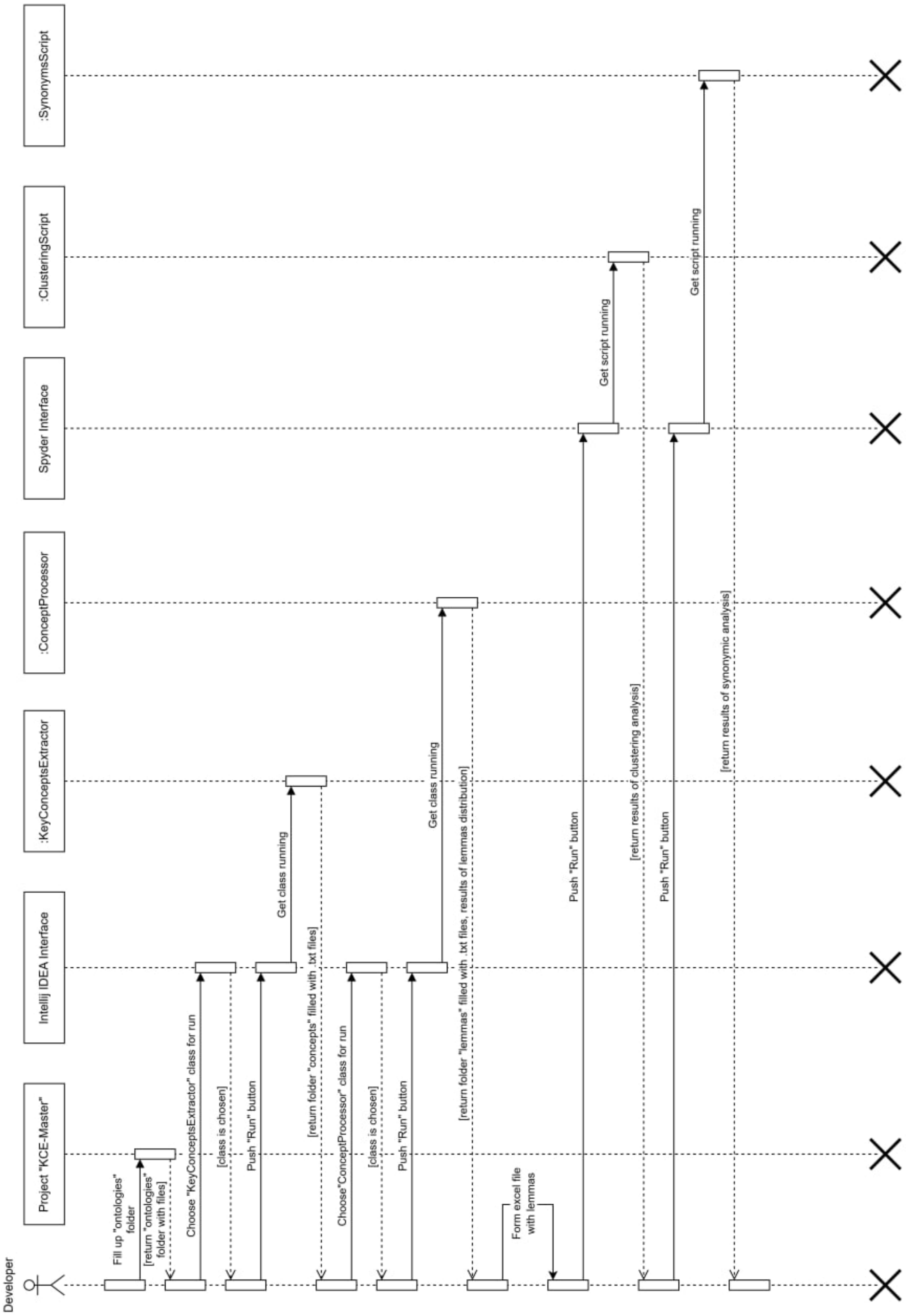
Developer
Project "KCE-Master"
Intellij IDEA Interface
:KeyConceptsExtractor
:ConceptProcessor
Spyder Interface
:ClusteringScript
:SynonymsScript
Fill up "ontologies" folder
[return "ontologies" folder with files]
Choose "KeyConceptsExtractor" class for run
[class is chosen]
Push "Run" button
Get class running
[return folder "concepts" filled with .txt files]
Choose"ConceptProcessor" class for run
[class is chosen]
Push "Run" button
Get class running
[return folder "lemmas" filled with .txt files, results of lemmas distribution]
Form excel file with lemmas
Push "Run" button
Get script running
[return results of clustering analysis]
Push "Run" button
Get script running
[return results of synonymic analysis]

Figure 5: Workflow Sequence Diagram of Semi-automated Analysis

5.1 Adjustments Overview

Key Concepts Extraction. The first step in concept processing involves placing all identified ontologies into the specific folder within the project directory to enable further analysis.

The application was reworked to automate the extraction of key concepts from many ontologies, rather than from a single predefined file. It was modified to iterate through all ontologies in specific folder (folder "ontologies"), apply the existing extraction algorithm [38], and save the results to individual text files.

Before being saved, the extracted key concepts are cleaned using a dedicated component. Previously, key concepts were represented by their full paths within the ontology. To improve interpretability, extracted concepts are cleaned by isolating their local names. It is done by taking only final fragments of their URIs. This transformation removes structural identifiers such as path delimiters. As a result, the cleaned output consists of human-readable terms, which is then used for further processing, such as lemmatization.

After the concepts are cleaned, they are saved into files. Each file gets a name, corresponding to the ontology number, which was given during manual analysis. This ensures that each ontology's concepts are written into a distinct file. If a file with the same name already exists, it is deleted and recreated. Each saved file receives the key concepts of corresponding ontology into them line-by-line. All the created files with concepts are saved into another specific folder (folder "concepts").

Lemmatization. A specific class was added to the project to transform the extracted concepts into lemmas. This class performs the second stage of the processing workflow, taking cleaned concept files from the "concepts" folder and preparing them for further analysis.

First, files from the "concepts" folder are read. Once the concepts are loaded, they are passed through a lemmatization procedure, which applies morphological normalization using Stanford CoreNLP [31].

During manual analysis, it was observed that certain terms were difficult to lemmatize correctly and required special handling. To address this, the algorithm checks whether a term belongs to a predefined set of exceptions (e.g.,R&B, EP). If the term is listed as an exception, it is returned unchanged.

Additionally, the analysis revealed that semantically identical concepts were sometimes written with formatting differences, such as the use of underscores. To ensure consistent output, all underscores are removed from the resulting lemma strings.

The resulting lemmatized terms are saved into a new set of files. Similarly to the last saving procedure, the original input file name from "concepts" folder and each value is a list of corresponding lemmas, are taken. For each entry, a new output file name is generated. These files are saved in the new folder ("lemmas" folder), which is automatically created if it does not already exist. Each lemma is written on a separate line, following the same approach previously applied for saving concepts.

Lemma Analysis. The next step in the process involves analyzing the distribution of lemmas across the set of processed ontology files. This is handled by the new method. The method takes as input a map, in which each key corresponds to a filename, and the value is the list of lemmatized concepts found in that file. It outputs a new map, where each lemma is associated with a set of filenames in which it appears. This creates a file-to-lemma mapping.

By this, the analysis of how broadly each lemma is distributed across the ontology files, is proceeded. To identify the most frequently occurring lemmas, another method was introduced. It calculates the percentage of files in which each lemma appears. For the purposes of this study, a lemma is considered frequent if it occurs in more than 50% of the files. For finding lemmas existing in both ontologies, persentage was lowered to 30%. The algorithm performs the necessary counting and outputs the results to the console, including the frequency and the list of files in which each lemma was found.

Cluster Analysis. One of the final steps of analysis is performed through clustering, which is applied to group ontology files based on the distribution of their lemmatized concepts. The clustering is performed using `KMeans`, vectorization using `TF-IDF`, and dimensionality reduction using `PCA`. The input for this analysis is an Excel file, where each column represents a set of lemmas extracted from an individual ontology.

After the lemmas were saved in the last step of lemma analysis, they were exported to an Excel file for further analysis. The first row contained the names of the files from which the lemmas were extracted, while the subsequent rows listed the lemmas themselves.

For clustering analysis file with lemmas is loaded into a DataFrame, with each column treated as a separate document. The documents are transformed into TF-IDF vectors using the TfidfVectorizer from "scikit-learn", with English stop words removed.

To determine the most appropriate number of clusters, two metrics are calculated across a range of values: inertia (elbow method) - sum of squared distances to centroids, and silhouette score - a measure of cluster cohesion and separation.

With the optimal cluster count defined, the files are clustered using KMeans. Principal Component Analysis (PCA) is then applied to reduce the TF-IDF vectors to two dimensions for visualization. The resulting clusters are plotted, with each cluster represented in a distinct color.

For interpretability, each cluster is examined in terms of which files it contains and which lemmas occur frequently across those files. A lemma is considered frequent if it appears in at least 50% of the documents within that cluster. The output includes file names and a ranked list of these frequent lemmas.

Synonymic Analysis. Synonymic similarities of concept usage are being analyzed across different domain ontologies. The new script utilizes the same Excel file as the clustering analysis, where each column contains lemmas extracted from a single ontology. However, in this case, the script performs a different operation. It includes the use of the WordNet lexical database, which allows the detection of potential synonyms among terms present in the ontologies.

The lemmas are first loaded into a DataFrame, with their frequency distribution analyzed separately for music and movie domains. This is needed for future results, to explore which domain uses concepts more often. For each unique lemma in the dataset, the script queries the WordNet database to retrieve its synonym sets (synsets). It then checks whether any of the synonyms are present within the same set of collected lemmas. If a synonym is found, it is paired with the original lemma. If synonym is not found, lemma is not taken into consideration. Output shows lemmas by their frequency of appearence in each domain, with their synonyms. First ten most frequently met lemmas are used for analysis.

5.2 Initial Code Overview

The initial code, which served as the basis for further research, was retrieved from an open-source GitHub repository [37]. The logic behind this code is discussed in detail in the related paper by the author [38].

The paper discusses the development of an algorithm for the extraction of key concepts from a given ontology. The method considers various aspects of ontologies. First, it applies natural categories, together with density and coverage criteria. Natural categories are described through name simplicity, favoring single-word labels, and basic level, measuring the centrality of a concept within the ontology. In addition, global and local density are introduced to assess how richly a concept is described either within the overall ontology or within its immediate neighborhood. Following the first evaluation, an additional criterion, popularity, was incorporated. Most commonly used terms are subsequently selected based on their popularity both locally and globally.

Initial code implements the described algorithm. The ontology is loaded into the system, the ontology file is being parsed and then hierarchical taxonomy structure is built.

The resulting taxonomy is stored in an object, representing the ontology as a navigable graph of classes and relationships. This structure serves as the main input to the algorithm engine.

After construction, the taxonomy is processed by the specific class, which executes the key concept extraction pipeline. The specific engine evaluates each concept node using measures such as name simplicity, centrality, density, and coverage. The most representative concepts are returned based on the final ranking.

The number of outputted key concepts is predefined and equals 20. However, it may vary depending on the size and scores within the ontology.

5.3 Summary

In this chapter, a process of ontology data analysis and its implementation as a semi-automated analysis pipeline was presented. The process began with collecting and manually reviewing ontologies, followed by the extraction and cleaning of key concepts using a modified algorithm. These concepts were then lemmatized and examined for consistency, with manual interventions applied throughout to address duplicates and formatting inconsistencies. Lemmas were analyzed for their frequency and distribution across files. After that clustering was performed using different tools, including TF-IDF vectorization, dimensionality reduction, and evaluation metrics to determine optimal groupings. Another script works on detection of synonymic lemmas accross ontologies. The workflow combined automated techniques with iterative manual check to ensure the quality and interpretability of the results.

6 Empirical Evaluation

This section presents the overall outcomes of the conducted work. Initially, the results are outlined in terms of their direct outputs. Following this, a more detailed examination is provided, first addressing the results related to music ontologies, and subsequently those concerning movie ontologies. The section concludes with a comparative analysis aimed at identifying potential conceptual connections between the ontologies developed by students across the two domains.

The project outcomes derive from a sequence of actions. Initially, files with lists containing the main concepts extracted from each ontology were produced. These were followed by files presenting the corresponding lemmas of these concepts for each ontology. The most significant output, however, is the console-based summary listing the lemmas that appeared in more than 50% of the ontologies, along with the percentage of their occurrence and the names of the corresponding files in which they appeared.

The script produced a different type of analytical output. It was employed to cluster similar concepts across the ontologies. To determine the appropriate number of clusters, the Elbow method and the Silhouette analysis were applied. Based on these analyses, a clustering algorithm was executed. The resulting output consists of identified clusters containing lemmas, alongside the list of files that were grouped into each cluster.

Another important component of the project was the manual analysis of the extracted lemmas. This part produced its own set of results, separate from those generated by the automated tools.

6.1 Analysis Results for Music Ontologies

The analysis of music ontologies showed several notable findings. Following a manual review, a total of 48 ontologies were included in the overall automatic analysis.

During the clustering analysis, an unexpected pattern was observed. While the majority of files clustered as expected, certain pairs of files were repeatedly grouped together. A manual inspection of the lemmas and their corresponding concepts and ontologies revealed that these cases were in fact duplicates. Upon reviewing the origin of the ontologies submitted by students as part of their coursework, it was concluded that these were not merely duplicates, but instances of copied work. As a result, the affected ontologies were excluded from the final dataset, and the total scope of analyzed ontologies was adjusted accordingly.

Determining the appropriate number of clusters presented a challenge.

However, based on the evaluation of the Elbow Diagram, it was concluded that five clusters would be the most suitable choice (see Figure 6).

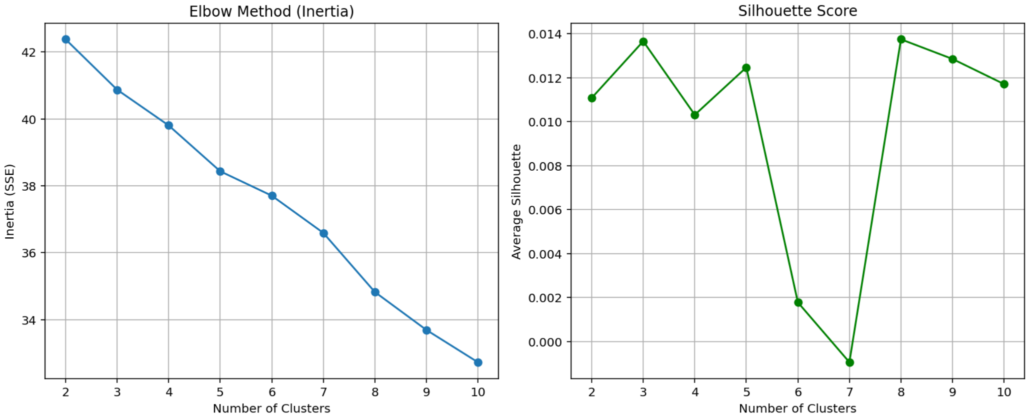

Figure 6: Elbow and Silouette Diagrams for Music Concepts

The results of the clustering analysis are presented in Figure 7. It can be observed that some clusters, such as Cluster 1 (green) and Cluster 4 (gray), are clearly distinguishable, while others, such as Clusters 0 (blue) and 2 (brown), exhibit notable similarity and spatial overlap. Overall, the clusters are moderately well-separated in the two-dimensional PCA space. This states that initially files include similar patterns in lemmas.

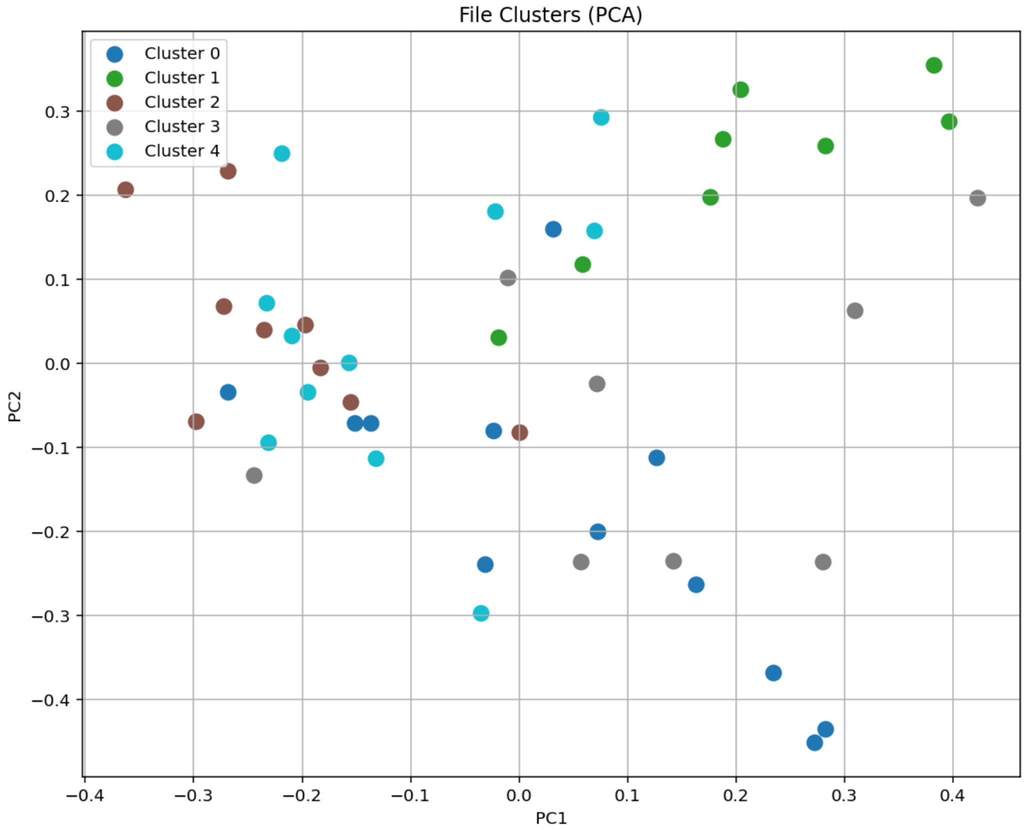

Figure 7: PCA for Music Concepts

The results of the most frequently met lemmas in clusters are shown in Table 1. From the table, several topic-specific patterns can be observed:

- **Cluster 0:** Focuses on artists - persons, albums, songs, and genres.
- **Cluster 1:** Emphasizes instruments, with lemmas such as *piano*, *drum*, *bass*, and *vocals*.
- **Cluster 2:** Focused on music production, including lemmas like *label* and *producer*.
- **Cluster 3:** Centers on music genres such as *jazz*, *blues*, *rock*, and *metal*.
- **Cluster 4:** Similar to Cluster 0, but emphasizes band artists and singers rather than individual artists.

Table 1: Frequent Lemmas by Cluster (Music Ontologies)

Cluster	Frequent Lemmas (count)	Files in Cluster
0	person (10), album (7), genre (7), song (7), artist (6)	12
1	album (7), song (7), instrument (7), guitar (7), genre (6), piano (6), bass (5), artist (4), band (4), drum (4), vocal (4)	8
2	genre (9), album (8), song (7), label (6), producer (5), person (5)	9
3	country (6), rock (6), genre (6), pop (5), artist (5), blues (4), jazz (4), album (4), metal (4), person (4), instrument (4)	8
4	album (7), band (7), singer (7), song (6), artist (6)	11

The results of lemmas comparison indicated that only a limited number of concepts were commonly used across the various music ontologies. The most frequently occurring lemmas included: genre, song, artist, person, and album. This pattern is also observed in the clustering results, where these lemmas appeared in nearly every cluster.

Table 2: Frequently Occurring Lemmas Across Music Ontologies

Lemma	Percentage of Files	Number of Files
genre	66.67%	36
album	66.67%	36
song	57.41%	31
artist	50.00%	27
person	50.00%	27

6.2 Analysis Results for Movies Ontologies

The results of movies ontologies were less extreme in comparison to music ones. A total of 62 ontologies were included in the overall automatic analysis following a manual review.

During the clustering analysis, no extreme outcomes were found, all of them were observed during the manual data analysis and failed ontologies were excluded from the final analysis.

With movies ontologies, finding the appropriate number of clusters in the analysis was a lesser challenge, as both Elbow and Silouette diagrams agreed upon four clusters (see Figure 8).

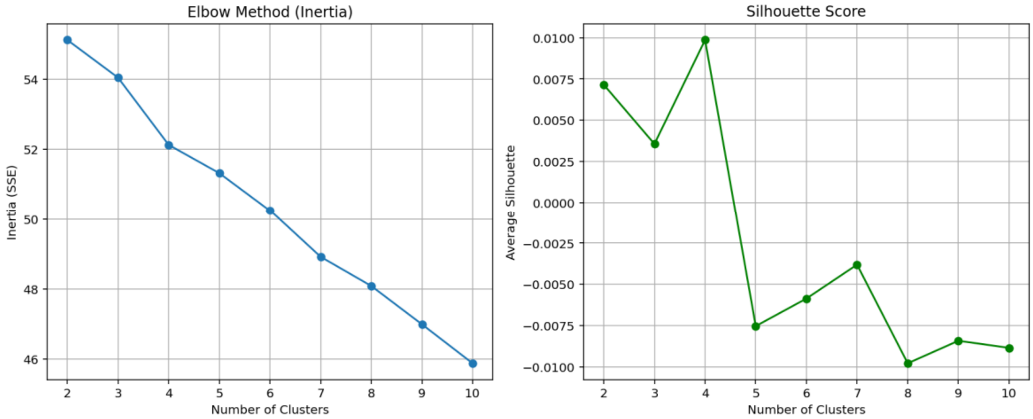

Figure 8: Elbow and Silouette Diagrams for Movies Concepts

The results of the clustering analysis are presented in Figure 9. The scatterplot shows the distribution of movie ontologies along the first two principal components (PC1 and PC2), based on their key concept features in the forms of lemmas. Four distinct clusters have been identified and are visualized with different colors: Cluster 0 (blue), Cluster 1 (red), Cluster 2 (pink), and Cluster 3 (cyan).

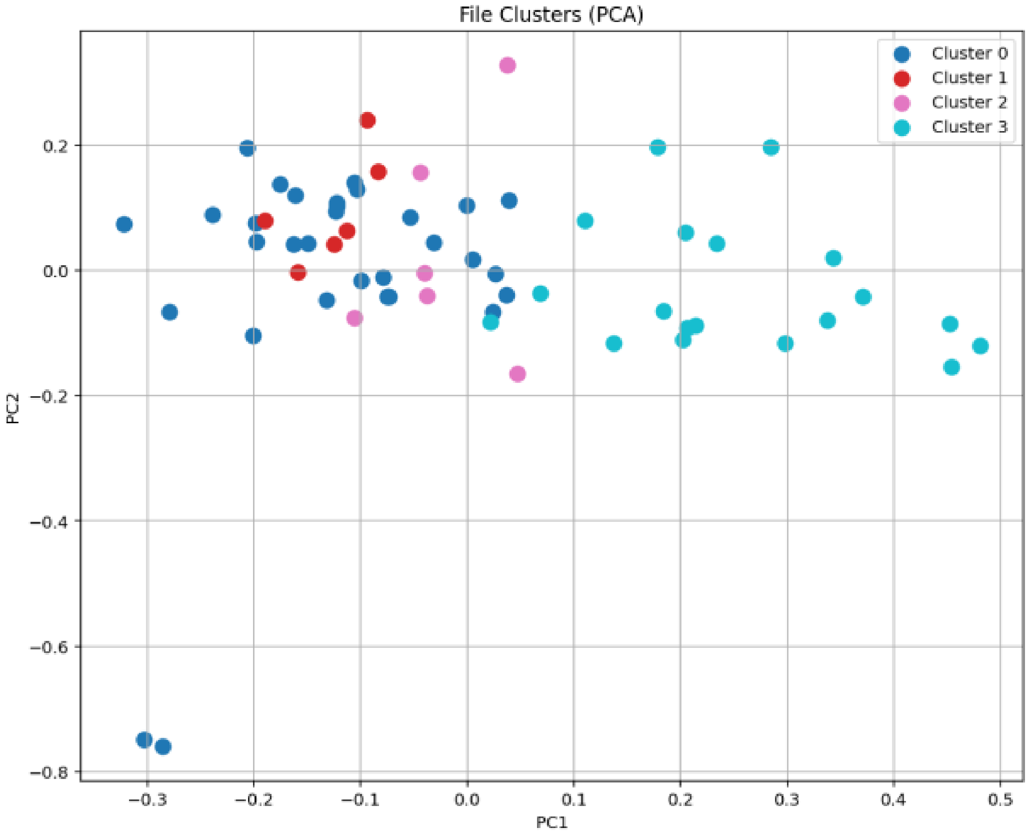

Figure 9: PCA for Movies Concepts

In contrast to the results for music ontologies, the clustering of movie ontologies exhibits a more structured and horizontally extended separation along the PC1 axis. Cluster 3 (cyan) is notably distinct, taking over the far-right region of the plot and demonstrating strong separation from the remaining groups. Cluster 0 (blue), which includes the largest number of files, is denser than others and centralized. This indicates a high level of conceptual similarity within this group. Clusters 1 (red) and 2 (pink) appear in closer proximity to Cluster 0 and show partial overlap with it. This suggests that ontologies within these clusters share several common concepts, despite being in different groups.

The results of the most frequently met lemmas in clusters are shown in Table 3. From the table, several topic-specific patterns can be observed:

- **Cluster 0:** Focused on the process of movie production and its outcomes, including terms such as *award*, *director*, and *producer*.

- **Cluster 1:** Emphasizes movie-making, with concepts related to roles in production such as *writer*, *composer*, and *editor*.

- **Cluster 2:** Centers around individuals, containing terms like *person*, *actor*, *writer*, *director*, and *producer*.
- **Cluster 3:** Genre-based, including movie genres such as *drama*, *thriller*, and *horror*.

Table 3: Frequent Lemmas by Cluster (Movie Ontologies)

Cluster	**Frequent Lemmas (count)**	**Files in Cluster**
0	movie (28), actor (27), director (26), person (24), award (21), producer (20), genre (19)	31
1	director (6), country (5), person (5), actor (5), writer (5), movie (4), producer (4), cast (3), company (3), award (3), composer (3), editor (3)	6
2	movie (5), person (5), actor (5), genre (5), director (4), writer (3), producer (3)	6
3	actor (19), comedy (17), genre (17), person (15), director (14), movie (14), action (12), horror (11), drama (10), thriller (10)	19

The results from the application showed that only a limited number of concepts were frequently used across the various movie ontologies. The most commonly occurring lemmas included: actor, movie, director, person, genre, and producer. This pattern is also reflected in the clustering results, where these lemmas appeared in the majority of clusters, indicating their central role in structuring domain knowledge.

Table 4: Frequently Occurring Lemmas Across Movie Ontologies

Lemma	**Percentage of Files**	**Number of Files**
actor	88.89%	56
movie	80.95%	51
director	79.37%	50
person	77.78%	49
genre	66.67%	42
producer	55.56%	35

6.3 Analysis Results for Connection between Different Domains

During the study, the discussion of overlapping concepts within ontologies was widely explored. For this reason, it was important to check possible connections between two ontologies of similar, creative topics. From the results, it was observed that only two concepts, "genre" and "person", overlap within those ontologies with a frequency of more than 50%, which is shown in Table 5. These concepts are quite abstract and can be used in various different domains beyond movies or music. Other overlapping concepts, such as "actor", "movie", "director", and "producer", are more specific. During further investigation, it was found that all of the terms except "album" are met in both ontology domains.

Table 5: Frequently Occurring Lemmas Across All Ontologies

Lemma	**Percentage of Files**	**Number of Files**	**Ontology Domain(s)**
genre	66.1%	78	Both
person	64.41%	76	Both
actor	48.31%	57	Both
movie	44.07%	52	Both
director	43.22%	51	Both
producer	38.14%	45	Both
album	30.51%	36	Music only

6.4 Analysis Results for Simplicity and Overlap

Through the analysis, it was found that both music and movie ontologies demonstrate domain-specific clusters of frequent lemmas. In both domains, the identification of the most commonly occurring terms was conducted. For the music domain, these included concepts such as *genre*, *album*, *song*, and *artist*. For the movie domain, the most frequent terms were *actor*, *movie*, *director*, and *producer*. Taken together, these terms are characteristic of their respective domains.

It was also observed that the most frequently occurring terms are typically short and cover broad areas of knowledge. In Table 6, we can observe how, as frequency decreases, the lexical forms of the terms become more complex. From short and cognitively simple terms, they shift to longer and more compound expressions, such as *songwriter* or *production company.*

Table 6: Lemma Occurrence by Domain and Frequency Slice

Domain	**Lemma**	**Number of Ontologies**
	Music Ontologies	
Most Frequent	album	36
	genre	36
	song	31
	person	27
	artist	27
Middle Frequent	instrument	17
	concert	14
	pop	14
	musician	13
	rock	13
Not Frequent	punk	5
	keyboard	5
	songwriter	5
	lyric	5
	publisher	5
	Movie Ontologies	
Most Frequent	actor	56
	movie	51
	director	50
	person	49
	genre	42
Middle Frequent	comedy	17
	rating	17
	location	14
	book	13
	language	13
Not Frequent	streamingplatform	5
	creativework	5
	city	5
	actress	5
	productioncompany	5

According to the results of the analysis, Table 7 shows the most frequently met concepts that have synonyms. The table states how many ontologies these terms are found in, per domain. The analysis used the WordNet library, which, as we can see from the results, is not always context-sensitive.

From the top result "person" we can see that the only synonym found is "soul," which, in the current context of music ontologies, refers to a genre of music rather than the human soul. This is a limitation of WordNet, which still exists even in widely used libraries. Similarly, this happened with the word "band", where the suggested synonym "set" lacks semantic relevance in this context. Therefore, critical analysis is always necessary when reviewing the results.

In the case of human-related terms, the word "director" was matched with both "manager" and "conductor," which, although technically correct in some contexts, are rarely used as synonyms in the analyzed domains. This shows how the same term may carry different meanings depending on domain-specific usage.

Interestingly, the word "writer" appears frequently in both domains. This is somewhat counter intuitive, as one would typically expect the term "author" to be associated with music or songs. This observation points to possible shifts in terminology usage.

It is also important to note that for certain terms, such as "artist," no synonyms were identified. This suggests that some domain-specific terms are not easily interchangeable and lack equivalents within the WordNet lexicon.

Table 7: Most Frequent Word–Synonym Pairs Across Ontologies

Lemma	**Ontology Domain**		**Synonym**	**Ontology Domain**	
	Music	**Movies**		**Music**	**Movies**
person	27	49	soul	3	0
movie	1	51	film	0	7
cinema	1	13	film	0	7
director	1	50	manager	2	1
director	1	50	conductor	1	1
song	31	1	vocal	5	0
author	0	2	writer	3	25
rating	5	17	place	1	2
rating	5	17	rat	0	1
band	21	1	set	0	1

6.5 Interpretation of Results

The observed results confirm that key concepts tend to form around central thematic ideas shared within domain-specific ontologies. In the case of music ontologies, clusters were mostly oriented around structural and content-related topics, such as genres, albums, and songs. These concepts appeared to be the most recurrent and indicate what students as creators of ontologies perceive them as essential in describing the domain.

For the movie ontologies, the clusters pointed toward individuals and their professional roles within the filmmaking process: *actor*, *director*, and *producer*. The consistent pattern in music ontology construction shows that it leans toward identifying people and their functions. This supports the theory of Barry Smith [50], pointing out that more frequently used concepts are the easier recognizable ones.

When comparing both domains, overlapping lemmas, such as *person* and *genre*, indicate that abstract, high-level concepts are used across different types of creative ontologies. These lemmas are likely perceived as "cognitively economical" and therefore are more frequently reused.

The table of synonym usage supports this by demonstrating that only a limited number of concepts appeared with meaningful variations across the ontologies. Most synonyms, such as *movie–cinema-film*, were found only a few times. This suggests that while variation in vocabulary exists, it is not widespread, and many ontologies rely on a shared core of terminology, especially within the same domain.

These patterns confirm the initial assumption that humans tend to use simple, abstract, and widely recognizable concepts to build ontologies. Moreover, the results validate that the semi-automated pipeline used in this study is efficient in revealing such conceptual structures and therefore can be reused for future research.

6.6 Prior Research Comparison

Current work is a novel approach in the analysis of key concepts. One of the most valuable previous studies regarding the topic of key concept extraction is the work of Silvio Peroni [37, 38]. As the current study implements the application code developed by Peroni, it therefore builds upon and extends it. The work of Peroni introduced valuable perspectives on how key concepts are defined within ontologies, expressing it through different angles such as simplicity, centrality, and frequency. His study allowed the current one to evolve and build further on the analytical foundation.

6.7 Summary

For the music ontologies, five clusters were identified. They focused of such themes as artists, instruments, music production, and genres. Manual inspection revealed duplicated submissions, which were excluded from the final analysis. Frequently recurring lemmas included *genre*, *album*, and *song.*

In the movie ontology analysis, four clusters emerged, each representing themes, such as production roles, individuals, and genres. High-frequency lemmas included *actor*, *movie*, and *director*.

A comparative analysis of the two domains identified overlapping concepts, notably *genre* and *person*, which appeared in over 50% of the ontologies in both domains. Other shared but domain-weighted terms, such as *actor*, *movie*, and *producer*, highlighted thematic parallels across creative domains.

7 Conclusion

This chapter concludes and summarizes the finding of the study. It outpoints results, connects them to the research questions stated in the introductory part of the paper. Further it discusses the limitations of work as well as possible solutions and remarks for future research,

7.1 Summary of Findings

The main goal of the current study lies in understanding the types and degrees of overlap among key concepts within ontologies. This was achieved through addressing the main research question and three subquestions:

RQ: *To what extent do student-authored ontologies overlap in terms of the key concepts they use?*

To answer this question, firstly a literature review was conducted, which demonstrated that the full picture cannot be understood without addressing the subquestions that reveal the core aspects of the main research question. In general, to analyze key concepts, a structured analytical process was developed for this study, supported by a custom application and a set of scripts. The application was built based on Peroni's work and was inspired by his study on key concept extraction [37, 38]. It follows three stages: determination of key concepts, lemmatization, and concept comparison. Each of these stages is essential for the subsequent data analysis, which supports answering the subquestions of this research.

For the analysis, two ontology domains were selected as the most frequently encountered - music and movies. A total of 55 music ontologies and 62 movie ontologies were used in the study, respectively.

The first subquestions to help find answer to the main one is:

SRQ1: *How does the degree of concept overlap vary across different ontology domains?*

Ontologies from both domains were processed through the application to determine the answer to this question. From the results, it was found that within the music domain, the most frequently occurring words are "genre" and "album", which appear in 66.67% of ontologies, followed by "song" at 57.41%, and "artist" and "person", each present in 50% of the cases. Further analysis using script for clustering revealed that these same lemmas form the largest cluster, which includes 12 ontologies. This suggests that within the same domain, most ontologies tend to express similar conceptual structures and vocabulary.

Results for ontologies in the movie domain showed similar results. The most frequently occurring lemmas are "actor" (88.89%), followed by "movie" (80.95%), "director" (79.37%), "person" (77.78%), "genre" (66.67%), and "producer" (55.56%). Similar to the music ontologies, the largest cluster for movies domain consists of 31 ontologies and is formed by most of these lemmas.

Overlap of key concepts between the two domains has also been identified. It revealed that the most frequently shared concepts are "genre" (66.1%) and "person" (64.41%).

Therefore, concepts can overlap not only within the same domain but also across different domains, which reflects how human cognition tends to organize and perceive information in similar ways.

To further deepen understanding of the forming process, the second subquestions has been answered:

SRQ2: *Are the overlapping key concepts predominantly general, high-level terms (hypernyms) with simple lexical forms (e.g., "person", "artist"), as opposed to more specific terms (e.g., "instrumental music")?*

From the analysis of key concepts, it was found that the most frequently occurring lemmas tend to be short words, typically not exceeding 5-6 letters. In contrast, concepts of medium frequency are generally longer, and less frequent concepts often consist of either long words or multi-word expressions. This pattern is observed across both domains and supports the theory of Eleanor Rosch [46], which suggests that humans tend to optimize cognitive effort by relying on abstract yet simple-to-understand concepts that efficiently explain meaning while conserving mental energy.

The last sub-question is necessary to fully understand the nature of overlapping concepts, not only from their lexical form, but also from their semantic content. This allows us to identify cases where concepts, although written differently, were intended to represent the same fundamental idea and therefore also represent a form of conceptual overlap:

SRQ3: *To what extent do overlapping key concepts share similar meanings, while expressed using different lexical forms (e.g., "musician" vs. "artist")?*

Through the analysis in the script, which uses a WordNet similarity module, it was determined that only a limited number of key concepts shared the same meaning while being written differently. The concept "film" was found in 7 ontologies, while "cinema" appeared in 13, and "movies" in 51. For music ontologies, little to no such cases were observed. For some key concepts, such as "person" (27) and "soul" (3), a negative connection was identified, as "soul" was most likely used to refer to a musical style rather than the human soul. A similar case was observed for "song" (31) and "vocal" (5), where "vocal" likely referred to a part of the song rather than representing the same concept.

Overall, the answer to the main research question has been received. Results of the study depicted that students with similar educational background indeed tend to engineer ontologies using same concepts. The study showed that concepts tend to overlap between ontologies and that humans usually perceive the construction of ontologies through small building blocks, short words, rather than large and complex ones, which tend not to overlap in meaning and are used more concisely.

7.2 Contributions and Implications

Current work shows deep insights from the theoretical and methodological perspectives, as well as implementation.

Theoretical Contribution. From the theoretical point of view, the current study supports several already established positions in cognitive sciences, as well as presents novel results. It provides thoughtful insights into conceptual overlaps within domain-specific ontologies. It highlights how more abstract-like concepts tend to be reused across ontologies engineered by different individuals. This reinforces the theory of Rosch [46], which suggests that people tend to use less cognitively demanding terminology — favoring general yet semantically rich concepts. In addition, the study acknowledges the importance of ontology quality and structural clarity as discussed by [52], and supports the idea of identifying meaningful semantic units within ontologies as proposed by [58].

Methodological Contribution. From the methodological point of view, the current study represents a novel continuation of Peroni's work [37, 38], expanding beyond the extraction of key concepts to include further analysis. The proposed semi-automated method involves not only the identification of key concepts, but also their lemmatization and multi-layered analysis to form a more complete picture of how and which concepts are used in ontologies. This approach can be further applied to the analysis of ontologies in other domains, potentially leading to new insights and results.

Practical Contribution. From the practical point of view, the methodology, as well as the results of the current study, can serve as a foundation for the work of researchers aiming to explore connections between ontologies. The study can be approached from different perspectives. For some, valuable insights for the development of core ontologies can be drawn from it, as it demonstrates the practical work of individuals who engineer ontologies. For others, the study can show a new perspective on the intellectual and cognitive aspects underlying the ontology creation process. In the context of such an important task for IT developers as ontology alignment, this work can also provide meaningful input and contribute to accelerating the knowledge-sharing process.

Implications. The uses of the implemented tool for key concept extraction and analysis can take multiple directions. First, the tool can be applied to different sets of ontologies, as it is not embedded in a single specific domain. This allows exploration of various spheres of knowledge, which may potentially lead to new insights. Therefore, the tool presents itself as a unique instrument for ontology assessment. From the results of the analysis, new theoretical foundations for improving interoperability among ontologies may emerge. For instance, potential new rules or patterns for standardization in domain or core ontology engineering could be developed. The identified shared vocabulary within specific domains may further lead to the direct construction of a core ontology, which can facilitate knowledge exchange and contribute to advancements in science and data transfer.

7.3 Limitations and Future Research Opportunities

Limited Ontology Diversity and Domain Coverage. One of the main limitations of the current work lies in the number of ontologies used for the analysis. The study would have benefited from a larger set of ontologies, as this could have revealed additional patterns in the ontology engineering process, for instance, connections related to synonymity. Only two domains were used in the project due to the limited availability of ontology files from other fields. While this still allowed for the identification of certain relationships, it was not sufficient to fully explore the principle of key concepts in the possible construction of core ontologies. The study would have also benefited from data in the domain of theatre or other art-related fields that cover a significant area of knowledge.

Data Quality Issues. The quality of the received data also revealed limitations and became one of the central issues of the work. This issue was not expected and was not discovered until a specific stage of the automated analysis, when it became clear that some ontologies duplicated each other, while others contained incorrect or incomplete data. These issues were not easily detectable through manual review, therefore further automation could have helped to both accelerate the process and prevent such problems from occurring.

Dependence on Manual Analysis. This leads to the next limitation - the role of manual analysis in the study. Manual reviewconstituted a significant part of the study. While it was essential for ensuring data accuracy and interpretation, it was also considerably time-demanding and required not only consistent effort but also a high level of attention to detail. Some of the verification steps could have been integrated into the processing pipeline to reduce the manual workload and streamline the workflow.

Lack of Pipeline Automation. Following this, the absence of a fully automated pipeline for the application and scripts stands out as another limitation. In the current setup, the developer is required to manually debug and run the code at each stage of the process, starting from key concept extraction and lemmatization to clustering and synonym detection. This design was chosen to allow full control and validation at each step. However, with additional checks and full automation, the overall analysis could become significantly faster and more efficient.

Limitations of Semantic Tools. The semantic modules used in this work also represent a limitation. For instance, the Stanford NLP module demonstrated restrictions in the range of lemmas it could extract from concepts, which required the use of a stop list, as some outputs became nonsensical after processing. Another used tool, the WordNet module, which was used for synonym detection, showed inaccuracies by linking words with unrelated meanings. These tools were selected due to their popularity and availability; however, further exploration and analysis of alternative modules could have led to better results.

Overall, future research may involve working with a larger dataset, as well as improvements in the codebase to enable a smoother and more automated workflow.

7.4 Concluding Remarks

The current study discovered the topic of overlap of key concepts in student-authored ontologies, focusing on music and movies domains. Application of semi-automed pipeline for analysis allowed the identification of recurring lemmas, extracted from concepts, analyse their synonymity, as well as detect thematic clusters and compare concepts within and across domains.

Findings of the study revealed that certain abstract and general concepts, such as *genre* and *person*, tend to appear consistently across different ontology domains, suggesting shared cognitive patterns in knowledge representation. At the same time, more domain-specific terms (e.g., *actor*, *album*, *movie*) also showed high frequencies, which showed practical needs of domain modeling.

The study further demonstrated how results support the theories of cognitive science on topics of cognitive economy and reusability. Creators of ontologies often rely on simple, short, and cognitively efficient terms. While some variation in terminology exists (e.g., through synonyms), overlaps are more often found in the use of core, high-level concepts.

Overall, the developed and tested approach provides valuable insights for ontology analysis and engineering. It offers a foundation for future work for improvement of ontology alignment, reuse, and interoperability across domains.

References

[1] Ontology search portal. https://ontology-search.ai.wu.ac.at/. Accessed: 2025-02-15.

[2] Harith Alani, Christopher Brewster, and Nigel Shadbolt. Ranking ontologies with aktiverank. In *The Semantic Web-ISWC 2006: 5th International Semantic Web Conference, ISWC 2006, Athens, GA, USA, November 5-9, 2006. Proceedings 5*, pages 1–15. Springer, 2006.

[3] Yoo Jung An, Kuo-chuan Huang, and James Geller. Naturalness of ontology concepts for rating aspects of the semantic web. *Communications of the IIMA*, 6(3):7, 2006.

[4] Bo Andersson and Vassil Momtchev. D7a. 1.1 larkc requirements summary and data repository. 2008.

[5] Fatima Ardjani, Djelloul Bouchiha, and Mimoun Malki. Ontology-alignment techniques: survey and analysis. *International Journal of Modern Education and Computer Science*, 7(11):67, 2015.

[6] Tim Berners-Lee, James Hendler, and Ora Lassila. Web semantic. *Scientific American*, 284(5):34–43, 2001.

[7] Pascal Boyer and H Clark Barrett. Domain specificity and intuitive ontology. *The handbook of evolutionary psychology*, pages 96–118, 2015.

[8] Anila Sahar Butt, Armin Haller, and Lexing Xie. Dwrank: Learning concept ranking for ontology search. *Semantic Web*, 7(4):447–461, 2016.

[9] Valentina Anita Carriero, Marilena Daquino, Aldo Gangemi, Andrea Giovanni Nuzzolese, Silvio Peroni, Valentina Presutti, and Francesca Tomasi. The landscape of ontology reuse approaches. In *Applications and practices in ontology design, extraction, and reasoning*, pages 21–38. IOS Press, 2020.

[10] Roger Chaffin and Arnold Glass. A comparison of hyponym and synonym decisions. *Journal of Psycholinguistic Research*, 19:265–280, 1990.

[11] Balakrishnan Chandrasekaran, John R Josephson, and V Richard Benjamins. What are ontologies, and why do we need them? *IEEE Intelligent Systems and their applications*, 14(1):20–26, 2002.

[12] Dhivya Chandrasekaran and Vijay Mago. Evolution of semantic similarity—a survey. *Acm Computing Surveys (Csur)*, 54(2):1–37, 2021.

[13] Stefan Decker, Sergey Melnik, Frank Van Harmelen, Dieter Fensel, Michel Klein, Jeen Broekstra, Michael Erdmann, and Ian Horrocks. The semantic web: The roles of xml and rdf. *IEEE Internet computing*, 4(5):63–73, 2000.

[14] Martin Doerr, Jane Hunter, and Carl Lagoze. Towards a core ontology for information integration. 2003.

[15] Teal Eich, David Parker, Yunglin Gazes, Qolamreza Razlighi, Christian Habeck, and Yaakov Stern. Towards an ontology of cognitive processes and their neural substrates: a structural equation modeling approach. *PloS one*, 15(2):e0228167, 2020.

[16] Mirna El Ghosh, Habib Abdulrab, Hala Naja, and Mohamad Khalil. Using the unified foundational ontology (ufo) for grounding legal domain ontologies. In *KEOD*, pages 219–225, 2017.

[17] Günes Erkan and Dragomir R Radev. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research*, 22:457–479, 2004.

[18] Miriam Fernández, Iván Cantador, Vanesa López, David Vallet, Pablo Castells, and Enrico Motta. Semantically enhanced information retrieval: An ontology-based approach. *Journal of Web Semantics*, 9(4):434–452, 2011.

[19] Asunción Gómez-Pérez. Ontology evaluation. In *Handbook on ontologies*, pages 251–273. Springer, 2004.

[20] Google Scholar. Google scholar. https://scholar.google.com/, 2025. Accessed: 2025-06-16.

[21] Michael Granitzer, Vedran Sabol, Kow Weng Onn, Dickson Lukose, and Klaus Tochtermann. Ontology alignment—a survey with focus on visually supported semi-automatic techniques. *Future Internet*, 2(3):238–258, 2010.

[22] Nicola Guarino. Semantic matching: Formal ontological distinctions for information organization, extraction, and integration. In *International Summer School on Information Extraction*, pages 139–170. Springer, 1997.

[23] Nicola Guarino. Understanding, building and using ontologies. *International journal of human-computer studies*, 46(2-3):293–310, 1997.

[24] Nicola Guarino. *Formal ontology in information systems: Proceedings of the first international conference (FOIS'98), June 6-8, Trento, Italy*, volume 46. IOS press, 1998.

[25] Nicola Guarino and Christopher Welty. Evaluating ontological decisions with ontoclean. *Communications of the ACM*, 45(2):61–65, 2002.

[26] Sathvik Guru Rao. Ontology matching using domain-specific knowledge and semantic similarity. Master's thesis, University of Twente, 2022.

[27] James A Hampton. Concepts as prototypes. *Psychology of learning and motivation*, 46:79–113, 2006.

[28] Sandra Heiler. Semantic interoperability. *ACM Computing Surveys (CSUR)*, 27(2):271–273, 1995.

[29] Matthew Horridge and Sean Bechhofer. The owl api: A java api for owl ontologies. *Semantic web*, 2(1):11–21, 2011.

[30] J Kwak and Hwan-Seung Yong. Ontology matching based on hypernym. *International journal of Web & Semantic Technology (IJWesT)*, 1(2), 2010.

[31] Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60, 2014.

[32] Joseph McCaffrey and Jessey Wright. 14 neuroscience and cognitive ontology: A case for pluralism. *Neuroscience and philosophy*, page 427, 2022.

[33] Eric Miller and Ralph Swick. An overview of w3c semantic web activity. *Bulletin of the American Society for Information Science and Technology*, 29(4):8–8, 2003.

[34] Fabian Neuhaus and Janna Hastings. Ontology development is consensus creation, not (merely) representation. *Applied Ontology*, 17(4):495–513, 2022.

[35] Ontotext. Astrazeneca boosted early hypotheses testing by using ontotext's lld inventory, 2023. Accessed: 2025-05-30.

[36] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[37] Silvio Peroni. Kce - key concepts extraction. https://github.com/cbobed/KCE, 2024. Accessed: 2025-01-07.

[38] Silvio Peroni, Enrico Motta, and Mathieu d'Aquin. Identifying key concepts in an ontology, through the integration of cognitive principles with statistical and topological measures. In *The Semantic Web: 3rd Asian Semantic Web Conference, ASWC 2008, Bangkok, Thailand, December 8-11, 2008. Proceedings. 3*, pages 242–256. Springer, 2008.

[39] Frederic Portoraro. Automated reasoning. 2001.

[40] María Poveda-Villalón, Asunción Gómez-Pérez, and Mari Carmen Suárez-Figueroa. Oops!(ontology pitfall scanner!): An on-line tool for ontology evaluation. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 10(2):7–34, 2014.

[41] Edson Prestes, Joel Luis Carbonera, Sandro Rama Fiorini, Vitor AM Jorge, Mara Abel, Raj Madhavan, Angela Locoro, Paulo Goncalves, Marcos E Barreto, Maki Habib, et al. Towards a core ontology for robotics and automation. *Robotics and Autonomous Systems*, 61(11):1193–1204, 2013.

[42] Shahzad Qaiser and Ramsha Ali. Text mining: use of tf-idf to examine the relevance of words to documents. *International journal of computer applications*, 181(1):25–29, 2018.

[43] Ripal Ranpara. A semantic and ontology-based framework for enhancing interoperability and automation in iot systems. *Discover Internet of Things*, 5(1):1–12, 2025.

[44] ResearchGate. Researchgate. https://www.researchgate.net/, 2025. Accessed: 2025-06-16.

[45] Lior Rokach and Oded Maimon. Clustering methods. *Data mining and knowledge discovery handbook*, pages 321–352, 2005.

[46] Eleanor Rosch. Principles of categorization. In *Cognition and categorization*, pages 27–48. Routledge, 2024.

[47] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

[48] Miriam Seoane Santos, Pedro Henriques Abreu, Nathalie Japkowicz, Alberto Fernández, and João Santos. A unifying view of class overlap and imbalance: Key concepts, multi-view panorama, and open avenues for research. *Information Fusion*, 89:228–253, 2023.

[49] Daniel Schober, Waclaw Kusnierczyk, Suzanna E Lewis, Jane Lomax, CJ Mungall, P Rocca-Serra, B Smith, and SA Sansone. Towards naming conventions for use in controlled vocabulary and ontology engineering. In *The 10th Annual Bio-Ontologies Meeting*, 2007.

[50] Barry Smith. Ontology. 2012.

[51] Chien DC Ta and Tuoi Phan Thi. Automatic evaluation of the computing domain ontology. In *Future Data and Security Engineering: Second International Conference, FDSE 2015, Ho Chi Minh City, Vietnam, November 23-25, 2015, Proceedings 2*, pages 285–295. Springer, 2015.

[52] Samir Tartir, I. Budak Arpinar, Michael Moore, Amit P. Sheth, and Boanerges Aleman-Meza. Ontoqa: Metric-based ontology quality analysis. In *Proceedings of the International Conference on Semantic Web and Databases (SWDB)*. Kno.e.sis Center, Wright State University, 2005.

[53] Mohammad Mustafa Taye. Understanding semantic web and ontologies: Theory and applications. *arXiv preprint arXiv:1006.4567*, 2010.

[54] Tippaya Thinsungnoena, Nuntawut Kaoungkub, Pongsakorn Durongdumronchaib, Kittisak Kerdprasopb, Nittaya Kerdprasopb, et al. The clustering validity with silhouette and sum of squared errors. *learning*, 3(7):44–51, 2015.

[55] Vienna University of Economics and Business. Wu library. https://www.wu.ac.at/en/library, 2025. Accessed: 2025-06-16.

[56] Robin Whittemore and Kathleen Knafl. The integrative review: updated methodology. *Journal of advanced nursing*, 52(5):546–553, 2005.

[57] World Wide Web Consortium (W3C). Rdf 1.1 primer. https://www.w3.org/TR/rdf11-primer/, 2014. Accessed: 2025-05-23.

[58] Gang Wu, Juanzi Li, Ling Feng, and Kehong Wang. Identifying potentially important concepts and relations in an ontology. In *International Semantic Web Conference*, pages 33–49. Springer, 2008.

[59] Xindong Wu, Vipin Kumar, J Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J McLachlan, Angus Ng, Bing Liu, Philip S Yu, et al. Top 10 algorithms in data mining. *Knowledge and information systems*, 14:1–37, 2008.