Dynamic knowledge graph evaluation

Roos Bakker [a,b,*] and Maaike de Boer [a]
[a] *Department Data Science, TNO, The Netherlands*
E-mails: roos.bakker@tno.nl, maaike.deboer@tno.nl
[b] *Leiden University Centre for Linguistics (LUCL), Leiden University, The Netherlands*

Abstract. In a world where information is exchanged at an increasing pace, knowledge becomes quickly outdated. Formal constructs that capture human knowledge, such as knowledge graphs and ontologies, need to be updated and evaluated to stay relevant and functioning. However, manually updating and evaluating existing knowledge models is labour intensive and prone to errors. This study addresses the challenge of evaluating changes in existing knowledge graphs. In this work, syntactic and semantic metrics tailored for change evaluation are introduced. The metrics are implemented and tested through experiments on knowledge graphs across various domains. In these experiments, real-world changes are simulated by removing concepts and introducing faulty ones before measuring the quality with the syntactic and semantic metrics. The hypothesis is that such changes decrease scores: removing concepts influences syntactic qualities such as the structure of the model, while adding faulty concepts affects semantic qualities like model consistency. The results confirm the hypothesis, showing that the extent and nature of the changes influence the scores. Additionally, size and degree of specialisation of the graph affect the scores. Overall, this study presents a set of evaluation metrics and provides empirical evidence of their efficacy in assessing modifications to knowledge graphs from different domains.

Keywords: Knowledge Graph Evaluation, Ontology Evaluation, Knowledge Graphs, Ontology Evolution, Change Evaluation Metrics

1. Introduction

Knowledge Graphs, ontologies, and similar constructs play an increasing role in our landscape of growing information. They structure and organise it, improve interoperability, and enable effective communication among diverse domains [61]. As these models evolve to accommodate changes and new information, the need for robust automatic evaluation metrics becomes urgent. This paper introduces a collection of evaluation metrics for changing knowledge graphs, and evaluates their effectiveness on (sub)graphs from multiple domains.

Previous work on knowledge graph or ontology evaluation has often relied on manual evaluation methods, the existence of a ground truth, or the usage within an application. While these approaches can ensure high quality, they are labour-intensive, prone to subjectivity, and not scalable for larger models. Furthermore, existing frameworks often focus on a static model and do not facilitate the tracking and assessing of changes. This is best illustrated with an example. Consider a part of a knowledge graph about occupations that describes skills of a *Data Scientist*, a *Data Analyst*, and a *Music Teacher*. This example can be viewed in a simplified graph form in Table 1. In this scenario, a knowledge graph developer needs to update the knowledge graph after the domain experts have agreed that the skill *perform data cleansing* should be added. By accident, the knowledge graph developer assigns the skill to the occupation *Music Teacher*. Now there is an error in the knowledge graph, which decreases its quality, and it can only be discovered by either the developer or a user who manually checks the contents. By automatically evaluating

*Corresponding author. E-mail: roos.bakker@tno.nl.

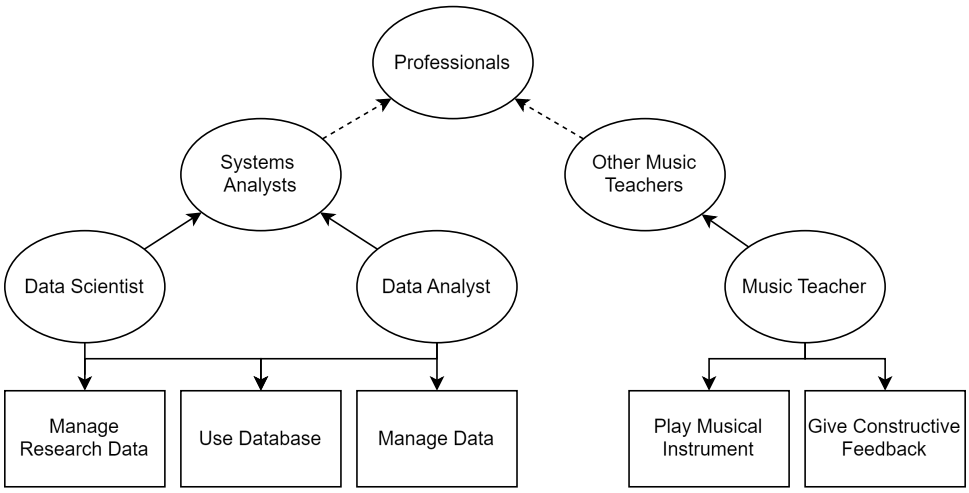

Fig. 1. Example graph

the knowledge graph, or part of it, score decreases can indicate errors and therefore the need to review the change. This paper proposes a set of metrics for the automatic evaluation of knowledge graphs on a semantic and a syntactic level, which can also be applied to smaller sections to better assess the impact of changes.

To validate the proposed metrics, an empirical analysis is conducted on three distinct knowledge graphs: the European Skills, Competences, Qualifications, and Occupations ontology (ESCO) [30], the Medical Subject Headings (MeSH) taxonomy [75], and the Pizza Ontology [103]. These models were chosen for their diverse applications, varying complexity, and widespread adoption within their respective domains. Specifically, the ESCO ontology is used within the labour market domain, where jobs and skills evolve with society, and thereby necessitating change evaluation. MeSH, on the other hand, is used for the medical domain, where ongoing innovations in medicine and treatments highlight the need for adaptive changes. Through this empirical analysis, the effectiveness and applicability of the metrics is demonstrated across different knowledge graph types, sizes, and domains.

This work extends the field of ontology and knowledge graph evaluation by introducing automatic evaluation metrics. Additionally, it provides insights into the performance of such metrics, and showcases the challenges that each unique knowledge graph poses. Furthermore, the empirical analysis shows that the quality of changes to a knowledge graph can be measured, thereby providing feedback to a knowledge graph developer and possibly prevent them from making mistakes.

In the next section, the field of ontology evaluation, evolution and existing tools is discussed. In Section 3, the evaluation metrics tailored for change evaluation are introduced. In Section 4, the different knowledge graphs are presented for testing the metrics and the experiment setting is discussed. Section 5 presents the results from the experiments. Finally, in Section 6 and 7 the results are discussed and interpreted, our findings are summarised, and future opportunities in this field are presented.

2. Related work

Ontologies and knowledge graphs are used to capture real-world knowledge of a domain in a formal way. They enable data interoperability, standardisation, and knowledge management in general. They contain at least a set of concepts and relations between them [53]. Stemming from early studies in philosophy, the study to ontology is as old as two and a half millennia, with the term ‘ontology’ following in the 17th century [91]. Computer science has been increasingly using the term since the 90s of the last century, where the meaning slowly shifted to a knowledge engineering approach [7, 46]. In this more technical viewpoint, several research directions have become more popular over the years such as ontology learning, ontology matching, and ontology evolution [5, 72, 113]. The term

knowledge graph appeared as early as 1970 [45], but gained popularity around 2012 with the upcoming of the Semantic Web and the search engines [40]. In this work the authors proposed the following definition: "A knowledge graph acquires and integrates information into an ontology and applies a reasoner to derive new knowledge" [40]. Other sources use the term knowledge graph indistinguishable with knowledge base, graph database, RDF graph and ontology. In current literature, the term knowledge graph has largely supplanted the term ontology, particularly in the fields of information retrieval, machine learning, and data science. In this work, we use knowledge graph as an umbrella term, including both ontologies and taxonomies. Given the focus of this work on evaluating dynamic knowledge graphs, the fields of ontology evaluation and ontology evolution will be discussed.

2.1. Ontology evaluation

Many approaches exist to evaluate a static (non-changing) ontology and its components. Wilson et al. [128] make a division between levels, scopes / aspects, techniques, stages, tools / methods and approaches / methodologies. This is slightly broader than for example the division of Lozano et al. [77], that takes into account dimensions, factors and characteristics or the division of McDaniel et al. [80] which makes the distinction between error checking, libraries, metric based, modularity and domain / task fit. Some examples are included from each category within the divisions of Wilson et al. [128].

Levels, or ontology layers, divide the evaluation based on levels by regarding an ontology as an multi-layered vocabulary. Dividino et al. [37] divide the assessment approaches into the syntactic (graph), semantic (conceptualisation) and pragmatic (communication) dimension making the ontology evaluation a semiotic triangle. Burton-Jones et al. [17] add social aspects to this. Other proposed levels include hierarchical aspects, lexicon, architecture, and context [128], and structural, functional and usability-profiling [48].

A different distinction can be made using scopes or aspects. These are structural intrinsic, domain intrinsic, domain extrinsic, application extrinsic [130]. These four scopes can be combined into internal quality and external quality, analogous to white-box and black-box testing in software development. These methods evaluate an application based on internal structure and external behaviour, respectively, and assess the influence of one on the other [131]. Additionally, the quality of the data and the schema can be separated for more precise evaluation.

Another angle is to look at the different phases or stages of the ontology in which the evaluation is done. Stages include analysis, design, implementation, deployment, maintenance. Degbelo [32] also makes a distinction between design and implementation stage. The conclusion of their work was that specific criteria such as accuracy, adaptability, and expressiveness can be used for evaluating the design of an ontology. Other criteria such as computational efficiency, practical usefulness, and precision and recall, can be used to evaluate ontologies at the implementation stage. From a different perspective, Bernabé et al. [11] add metrics on stability and goodness, and mainly focus on clustering for the ontology structure.

Yet another way to categorise evaluation is by looking at the techniques that are used for evaluation. Techniques include data-driven, human-based, golden standard based, application based (sometimes described as 'black box' approach [59, 80]), as also described in [13]. Hlomani et al. [59] match the techniques with the different levels, in which some levels are easier to approach evaluation that others. For example, on the level named 'context, application' the gold standard and data-driven technique is not applicable.

Besides high-over distinctions such as in levels or scopes, more specialist evaluation distinctions exists, such as using criteria or approaches. Characteristics, or criteria, are for example consistency, coverage, conciseness, correctness and modularity [128]. In an earlier paper, Wilson et al. divide the characteristics (then named criteria) into syntactic correctness, relevance, cognitive complexity, conciseness, consistency, adaptability, applicability, modularity, accuracy, completeness, efficiency, understandability, usability, and accessibility [129]. In their later paper [131] they add recoverability, timeliness, credibility, coverage, comprehensibility and compliance. Hlomani et al. [59] make a distinction between ontology quality and ontology correctness views on ontology evaluation. To measure ontology quality, they look at computational efficiency, clarity, and adaptability. To check the ontology correctness, they use accuracy, conciseness, consistency, and completeness.

Evaluation approaches or methodologies include ROMEO, GQM and OntoMetric. Requirements-Oriented Methodology for Evaluating Ontologies (ROMEO) is a methodology that helps ontology engineers determine appropriate methods for their ontology evaluation [137]. The methodology consists of the steps 1) role of the ontology,

2) ontology requirement, 3) question and 4) measure. Goal Question Metric (GQM) has the following steps: 1) establish goals, 2) develop questions; 3) establish data metrics; 4) design and test collection form; 5) collect and validate data and 6) analyse data [9]. Ontometric can also be seen as a methodology [77]. Guéret et al. [54] describe a framework for evaluating a Web of Data (LINK-QA), using the steps select, construct, extend, analyse and compare. Metrics include centrality, clustering, degree, description and same-as. Recently, Tsaneva [123] proposes an evaluation following the design science methodology focusing on a human-centric evaluation.

In this section, different distinctions are summarised in the evaluation metrics, such as levels, scopes, phases and techniques, but also criteria and approaches. Many tools and methods implementing these are available, as described in the next section.

2.2. Ontology evaluation tools

Several ontology evaluation tools exist, as well as some overview papers that compare them. Reiz & Sandkuhl [106] show that several older tools, such as Swoop, OntoQA, ODEval, OntoKBEval and OntoKeeper are not (publicly) available any more. McDaniel et al. [81] provide an overview of attributes for ontology quality assessment and the previous efforts from the literature. Recently, Gyrard et al. [55] provided an overview of frameworks and tools, the publication year, the specific topic and whether there is a prototype URL available. Topics include ontology quality survey, data quality survey / method, metrics, design patterns, ranking method and usability survey. The tools that are publicly available and focus on automatic ontology quality evaluation are OntoMetric [77], DOORS [81] and OQuaRE [39].

OntoMetric [77] uses 5 type of metrics: base metrics to measure the number of ontology elements (count of axioms, individuals, property links), schema metrics to analyse the design of the ontology (attribute richness, class-axioms ratio), graph metrics to analyse the taxonomy tree of the ontology (depth, absolute root cardinality), knowledge base metric to analyse the individuals and ontology population (average population, class richness, number of leaf classes) and class metrics to evaluate single classes (class readability, class inheritance richness). NEOntometrics [105] is the updated version of OntoMetrics, adding an interface and more efficient analysis of large ontologies and the calculation of evolutional metrics.

McDaniel et al. [81] introduced the DOORS system, a ranking framework for ontologies by using syntactic, pragmatic, semantic and social quality metrics. Each of these metrics is divided into several other metrics based on previous literature. These metrics include lawfulness, richness and structure for the syntactic quality, consistency, interpretability and precision for the semantic quality, accuracy, adaptability, comprehensiveness, ease of use and relevancy for the pragmatic quality and authority, history and recognition for the social quality.

OQuaRe [39, 121] does not attempt to create novel evaluation metrics, but combines metrics from literature in one tool. Current metrics include lack of cohesion, weight method count, depth of inheritance tree, number of ancestor classes, children and properties, response for a class, property richness, attribute richness, relations per class, inheritance relation richness and annotation richness.

A recent paper proposes Delta, which is a modular ontology evaluation system [71]. They use ontology statistics (number of classes, individuals, object properties, data-type properties and axioms), ontology metrics based on Lourdusamy & John [76] (size, appropriateness of module size, attribute richness, class/relation ratio, average population, equivalence ratio) and pitfalls (missing annotations, creating unconnected ontology elements).

Iyer et al. [63] create a SynEvaluator using rules to detect pitfalls and violations. They also implement a SemValidator with which crowd sourced surveys can validate the semantics of an ontology.

Heist et al. [58] create the framework KGrEaT to evaluate the performance impact of knowledge graphs on multiple downstream tasks, such as classification, regression and clustering.

In conclusion, many valuable tools exist, taking different angles, but all tools focus on static ontology evaluation. In the next section, the field of ontology evolution will be discussed and the evaluation in that field.

2.3. Ontology evolution and evaluation

Ontology evolution is a rising topic which focuses on the maintenance of ontologies. With more companies and institutions adopting semantic web techniques, the question arises how the ontologies can be kept up-to-date. Zablith

et al.[139] take a process-centric view on the topic, and introduce the ontology evolution cycle. The cycle contains five phases: the detection phase, a change suggestion phase, a change validation phase, an evolution impact phase, and lastly, a change management phase. The first phase, detection, detects whether there is need for a change in the ontology [21, 138]. In other literature this is also proposed as change capturing [118], or in the work of Zablith et al. [138], it is called information discovery. The second phase suggests changes to the ontology based on either text corpora or on other ontologies. Other corresponding terms are representation phase [118] and data transformation [68]. The change validation phase evaluates the changes from the previous step. This can be a domain focused validation, a formal validation, or a user validation. This stage is also present in most other works on ontology evolution. The evolution impact phase measures the impact of the changes, which is often done on an application level. This phase is also included in the work of Stojanovic et al.[118]. Finally, the change management phase records and versions changes in the ontology continuously. In this paper, the focus lies on on measuring the effect of changes on the quality of the ontology, which fits best in the change evaluation phase as proposed by Zablith et al. [139].

Proposed changes to the existing ontology should be evaluated to prevent inconsistencies and mismatches. Reiz et al. [104] show that the ontology metrics are a good foundation for the interpretation of the quality of an ontology, but that the choice of metrics is essential and current metrics need human interpretation of the results.

Changes can be evaluated with different methods, in which a distinction is often made between three levels: 1) domain properties-based or semantic [127], 2) formal properties-based or syntactic, and finally on impact according to [139]. In this work, the focus lies on the first two, which will be discussed in the next section.

3. Evaluation metrics

In this study, novel evaluation metrics are introduced that are designed to assess the quality of evolving ontologies and knowledge graphs, expanding upon established methodologies. This work connects to ontology evolution and specifically the change evaluation phase proposed by Zablith et al. [139]. Additionally, inspiration is drawn from the extensive literature on static ontology evaluation. The division of metrics is made on the syntactic and semantic evaluation framework outlined by McDaniel et al. [80] and mentioned by Zablith et al. [139], which itself builds upon several works in the field [17, 114, 119, 135]. Furthermore, a data-drive approach is adopted to allow for automatic evaluation of a knowledge graph. Considering the stages described in section 2, this work falls under the evaluation of the internal quality of a knowledge graph, or a white-box approach. The syntactic and semantic metrics are introduced in sections 3.1 and 3.2.

3.1. Syntactic quality metrics

The Syntactic Quality metrics evaluate the knowledge graph based on formal properties of the knowledge graph. These metrics measures aspects of the knowledge graph such as inheritance relationships, rules, and the richness of relations. McDaniel et al. [81] proposed several metrics for measuring these aspects, and this type of evaluation is also mentioned by Zablith et al. [139] to ensure that changes in a non-static knowledge graph will not invalidate constraints. In this work, the syntactic metrics for static ontology evaluation as proposed by McDaniel et al. [81] are extended and adapted to allow for change evaluation. The Syntactic Richness metric, which calculates the average number of attributes per class, is modified. To prevent this score from resulting in large, unbound scores, a normalisation step is introduced. Specifically, the average number of attributes per class (a) is normalised using an arctan function [34]. This normalisation ensures that the score of a lies between 0 and 1, instead of ranging from 0 to infinity in a knowledge graph with a theoretically infinite number of attributes. While the existing metrics from previous work are repeated for completeness, they have not been experimentally validated in previous work, an experimental validation will be included in section 4. The Syntactic Quality metrics are detailed in Table 1.

The metrics are repeated for completeness and since they have not been validated experimentally in previous work, they will be included in section 4. The Syntactic Quality metrics are described in Table 1. The Syntactic Lawfulness (SyL) measures the ratio of breached rules to the total set of rules. The Syntactic Richness (SyR) includes two scores: 1. The normalised average number of attributes per class, and 2. the ratio of non-inheritance

Table 1
Syntactic quality metrics

Metric	Calculation
Syntactic Quality (SyQ)	$SyQ = w_{s1} * SyL + w_{s2} * SyR + w_{s3} * SyS$
Syntactic Lawfulness (SyL)	Let b be the total number of breached rules in the knowledge graph. Let s be the number of statements in the knowledge graph. Then $SyL = b/s$
Syntactic Richness (SyR)	Let a be the average number of attributes per class in the knowledge graph. Let r be the ratio of non-inheritance relations to all relations in the knowledge graph. Then $SyR = w_{syr1} * \left(\frac{\arctan(a)}{\pi} + \frac{1}{2}\right) + w_{syr2} * r$.
Syntactic Structure (SyS)	Let s be the number of subclasses in the knowledge graph. Let c be the total number of classes in the knowledge graph. Then $SyS = s/c$.

relations to all relations. To balance these two scores, two weights are included (w_syr1, w_syr2), which can be given different weights depending on what aspect is more important in a certain knowledge graph. The Syntactic Structure (SyS) measures the ratio between subclasses and total classes. Finally, the Syntactic Quality (SyQ) is the weighted average of these scores, where the weights (w_sn) can be adjusted depending on the use case.

3.2. Semantic quality metrics

The Semantic Quality metrics evaluate the knowledge graph based on the information about the domain. These metrics measure the relevance of concepts and relations to the knowledge graph such as whether new terms are consistent with the existing terms, and whether terms in the knowledge graph are complete regarding to terms that are linked to the domain in other sources. The novel set of metrics shown in Table 2, which is inspired previous work [81], and designed to allow for automatic evaluation of changing knowledge graphs.

The Semantic Quality is the weighted average of the individual scores from the Semantic Consistency, the Semantic Interpretability, and the Semantic F1. Consistency is defined in [81] as the number of terms with inconsistent meanings divided by all the terms. This is clarified as the number of terms with semantic conflicts. Unfortunately, it does not become clear how these inconsistent meaning or semantic conflicts should be determined. Therefore we introduce an alternative metric. The Semantic Consistency (SeC) metric from Table 2 is introduced to measure the consistency within leaf groups. This is done by calculating the average similarity between concepts. The motivation for this is that concepts that are subclasses of the same node should be relatively similar. For instance, in the previous example in Table 1, the skills *manage research data* and *use database* are relatively similar. When a new skill is added that does not belong to the *Data Scientist* occupation, such as *play musical instruments*, the score lowers since it is not similar to the other skills. The Semantic Consistency can be measured locally or over the complete knowledge graph.

The Semantic Interpretability formula (SeI) in Table 2 is the same as McDaniel et al. [81] propose, however, practical implementation details, such as defining what constitutes an independent authority, are not provided. In this work, Wikipedia [125] is used as the independent authority. More concretely, terms in the knowledge graph with a Wikipedia page are divided by the total number of terms. With this implementation, the independent authority can easily be swapped for another database depending on the domain.

Semantic F1 (SeF) is similar to Semantic Interpretability on first sight. The difference is that Semantic Interpretability is defined as whether terms can be confirmed as a valid concept by an outside source, whereas the Semantic F1 score is about whether terms that should be in the knowledge graph are actually in there. McDaniel et al. [81] suggest precision as a measure, where they assess how ambiguous definitions are. Ideally, the knowledge graph should only have unambiguous definitions, thus have only precise terms. They suggest counting the number of definitions Wordnet[83] has on a concept, where having more definitions indicates more meanings, which in turn indicates an ambiguous term. This approach assumes that more definitions mean less precision, but this could also be the case for well known and widely used terms such as ‘data’ or ‘information’. One could argue that even though these terms have multiple interpretations, they are still essential for some domains, for instance in the context of a data scientist. Based on these insights, a more elaborate approach is taken in this work, where terms are gathered that are deemed essential to the domain, and the precision and recall are both taken into account in the Semantic F1 score. In this work, keywords are manually gathered from different sources. A keyword extraction algorithm

Table 2
Semantic quality metrics

Metric	Calculation
Semantic Quality (SeQ)	$SeQ = w_{e1} \cdot SeC + w_{e2} \cdot SeI + w_{e3} \cdot SeF$
Semantic Consistency (SeC)	Let $\{v_1, v_2, \ldots, v_n\}$ represent the vector representations of the individual concepts in the group, where each v_i is a vector in an n-dimensional space. The cosine similarity between two vectors v_i and v_j can be denoted as $\cos(\theta_{ij})$, where θ_{ij} represents the angle between the vectors. $$\cos(\theta_{ij}) = \frac{v_i \cdot v_j}{\|v_i\|\|v_j\|}$$ Let *LG* denote a Leaf Group with concepts: $\{v_1, \ldots, v_n\}$, where $n = \|LG\|$. The average cosine similarity (ACS) of the group can then be expressed as the sum of all pairwise cosine similarities divided by the total number of unique pairs: $$ACS(LG) = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \cos(\theta_{ij})$$ Finally, denote G as the set of leaf groups $\{lg_1, \ldots, lg_n\}$, where $n = \|G\|$, and ACS_i as the average cosine similarity score of the i-th group. $$SeC = \frac{1}{n} \sum_{i=1}^{n} ACS(lg_i)$$
Semantic Interpretability (SeI)	Let t be the total number of terms used to define classes and properties in the knowledge graph. Let w be the number of terms that have a sense listed in an independent authority. Then $SeI = w/t$.
Semantic F1 (SeF)	Let t be the total number of terms used to define classes and properties in the knowledge graph. Let d be the total number of definitions in an independent authority about the domain. Let c be the number of definitions that also occur in the knowledge graph. Then: $$precision = \frac{c}{d} \qquad recall = \frac{c}{t} \qquad SeF = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

is implemented using KeyBERT [52] for future automation of the process. However, for the experiments that are described in the next section, the manually gathered keywords were used to ensure their quality.

4. Validation

Work on metric-based evaluation of changing ontologies and knowledge graphs is scarce, as shown in Section 2. However, several tools and metrics exist for measuring the quality of static ontologies. This work expandec on such metrics and introduced new metrics in the previous section. In this section, the metrics are validated in experiments on three knowledge graphs from different domains: 1. the ESCO ontology for occupations and skills [30], 2. the MeSH taxonomy on diseases [75], and 3. the Pizza Ontology [102, 103]. The example graph from Table 1 is also included. These knowledge graphs were selected because the first two are actively used and maintained in real-world applications, while the last one is considered a gold standard in the ontology research community. This diverse selection allowed us to validate the metrics across a range of knowledge graph sizes and usage scenarios. The larger ESCO and MeSH ontology is split up into smaller subsets to test how the metrics work on different sizes of knowledge graphs.

4.1. Example graph

The example graph contains 5 nodes and 10 attributes: 2 occupation groups, 3 occupations, 5 skills, and 5 alternative labels for the skills. The example graph is visualised in the introduction in Table 1. This graph is meant as a simple example for the metrics. Three occupations from the ESCO ontology [30] are used: *Data Scientist* (code 2511.4), *Data Analyst* (code 2511.3) and *Music Teacher* (code 2354.1). For each occupation two skills were chosen, of which the skill *use databases* overlaps between *Data Scientist* and *Data Analyst*. *Data scientist* and *Data Analyst* are subclasses of the occupation group *Systems Analysts*, *Music Teacher* is a subclass of other music teachers. Finally, *Systems Analysts* and *Other Music Teachers* are indirect subclasses of the broad category professionals. Other superclasses are omitted for simplicity.

4.2. Pizza ontology

The pizza ontology models pizzas and their ingredients, and is considered a golden standard for ontology modelling [102, 103]. For applying the evaluation metrics, two different categories of pizzas were selected: meaty pizzas and vegetarian pizzas. Within these categories the *Giardinieri Pizza*, *Four Seasons* or *Quatro Stagioni*, and *Capricciosa* were selected. In the pizza ontology, the main attributes of the pizza are the toppings. These were included in this subgraph together with their alternative labels. This resulted in 3 nodes (the pizzas), and 86 attributes, of which 22 are toppings, and 64 alternative labels.

4.3. ESCO

ESCO, or the ontology of the taxonomy European Skills, Competences, Qualifications, and Occupations, is a classification and multilingual database system developed by the European Commission between 2011 and 2017 [30]. It serves as a common language to describe skills and competences in the labour market, and aims to facilitate better communication and understanding between different countries and sectors. ESCO describes occupations, skills, and qualifications, and can be used by employment services, education institutions, and employers in Europe [30]. The ontology aims to bridge the gap between education and employment by creating a shared vocabulary that can enhance the flexibility of workers and helps individuals make informed decisions about their careers. In this work, the ESCO ontology is split in three different sized subsets. The Example Graph described in Section 1 is also inspired by ESCO.

ESCO Data Scientist Subgraph The ESCO Data Scientist subgraph has 2 nodes and 196 attributes: The occupation group *Systems Analysts*, its subclass *Data Scientist*, and all data scientist skills (47) and alternative labels for the skills (149).

ESCO Data Scientist and Data Analyst Subgraph The ESCO Data Scientist and Data Analyst subgraph contains 3 nodes and 305 attributes. The nodes are the occupation group *Systems Analysts*, and two of its occupations *Data Scientist* and *Data Analyst*. The attributes are all of their skills, and alternative labels for the skills.

ESCO Systems Analysts Subgraph The ESCO Systems Analyst subgraph, the largest of the three, has 21 nodes and 2562 attributes. The occupation group *Systems Analysts* has 20 subclasses such as *Data Scientist* and *ICT Consultant*. All of their skills and alternative labels are included.

4.4. MeSH

MeSH is a controlled vocabulary taxonomy used for indexing and searching articles in the biomedical and health domain [75]. It is developed by the National Library of Medicine (NLM) and is used in the PubMed database. MeSH terms provide a standardised way to describe and categorise information related to diseases, anatomy, chemicals, and other biomedical concepts. The terms are categorised as a taxonomy. Researchers and healthcare professionals can use MeSH to improve the accuracy and efficiency of literature searches. This taxonomy provides an interesting use case due to the need for updates, even though it lacks the complexity in relations compared to an ontology. The medical domain is continuously updated with new research and insights, thus the need for evaluation of the changes and the taxonomy as a whole. For demonstrating and testing the metrics, two subsets of MeSH were created.

Table 3
Experiment 1: less labels for the example graph

Class	Attribute	Alternative Label
data scientist	manage research data	none
data scientist	use database	none
data analyst	use database	none
data analyst	manage data	none
music teacher	play musical instruments	none
music teacher	give constructive feedback	none

Table 4
Experiment 2: wrong labels for the example graph

Class	Attribute	Alternative Label
data scientist	manage research data	none
data scientist	use database	produce music from musical instruments
data analyst	use database	produce music from musical instruments
data analyst	manage data	manage data lifecycle
music teacher	play musical instruments	use database software
music teacher	give constructive feedback	propose constructive feedback

Diabetes Subgraph The Diabetes subgraph is extracted from the MeSH API using a SPARQL query on the latest version (d.d. August 2023), and contains 6 nodes and 2000 attributes. The Diabetes subgraph contains the items of the tree of *Diabetes Mellitus* (code C19.246), their attributes (meshv:allowableQualifier), and their alternative labels (rdfs:label).

Endocrine Subgraph The Endocrine subgraph is composed in the same way as the Diabetes subgraph, but now using the code for *Endocrine System Diseases* (C19). It has 9 nodes, and 2000 attributes, since only the first 1000 attributes for this subgraph were used.

4.5. Experimental setup

The goal of the experiments is to validate the metrics for measuring changes in three different knowledge graphs. In experiment 1, the labels of nodes are removed to make the knowledge graph more scarce, thereby testing the syntactic metrics. For the example graph, the resulting graph from experiment 1 is shown (in table format) in Table 3. In experiment 2, the semantic metrics are tested by changing alternative labels of nodes. Again, for the example graph this is shown in Table 4. Both experiments are compared to the original knowledge graph where no changes are made. We hypothesise the following:

Hypothesis 1 Removing information from the knowledge graph decreases the syntactic scores. Since the syntactic metrics should measure the structure (ratio of classes to subclasses) and the richness (number of attributes per class and ratio of non-inheritance relations to all relations), removing nodes will affect affect the Syntactic Richness metric, resulting in a lower Syntactic Quality metric.

Hypothesis 2 Changing labels such that the leaf nodes have wrong labels decreases the semantic scores. Having wrong labels under other classes mainly affects the Semantic Consistency metric, where the consistency of leaf groups is measured. In a carefully curated knowledge graph, leafs should be similar. By changing these leaf labels, the Semantic Consistency score will lower, and thus the Semantic Quality metric.

4.6. Implementation

The experimental setup was implemented in Python. The metrics were implemented as described in Section 3, the weights were distributed equally. For the Syntactic Richness, a power transformation of 20 was applied too all

Table 5
Validation results on the example graph

Exp.	Syntactic Validation			Semantic Validation			
	SyQ	SyR	SyS	SeQ	SeC	SeF	SeI
GT	**0.492**	0.384	0.6	**0.392**	0.754	0.172	0.25
Exp.1	0.462	0.324	0.6	0.274	0.336	0.111	0.375
Exp.2	**0.492**	0.384	0.6	0.352	0.633	0.172	0.25

results, since they were negatively skewed due to the large amounts of attributes in the subgraphs. For the Semantic Consistency, BERT [35] embeddings were used to embed concepts and calculate the cosine similarity between them. Similar to the Syntactic Richness, a power transformation of 10 is applied for the example graph and of 40 for the larger subgraphs to be able to show the differences in the results. When working with knowledge graphs with more diversity, this might not be necessary.

For the Semantic Interpretability, the Wikipedia API was queried to check whether a term had a page. Finally, for the Semantic F1, relevant keywords were manually gathered for the domain. For the ESCO subgraphs, keywords from the ResumeWorded website [107] were selected. For the MESH subgraphs, keywords were selected from diabetes.org and endocrine.org [3, 120]. Finally, for the pizza subgraph, no official organisation or list of keywords is available, therefore a list of keywords based on the Wikipedia pages on the relevant pizzas was created.[1]

5. Results

5.1. Example graph

For the example graph, the results of the syntactic and semantic metrics validation can be found in Table 5. The Syntactic Lawfulness cannot be measured due to the ESCO, MeSH, and pizza knowledge graphs not having rules, so the performance is not measured. The Syntactic Richness shows a lower score for experiment 1—where alternative labels are removed from the knowledge graph—compared to the ground truth and experiment 2. The Syntactic Structure score is the same among all experiments, because the structure is not changed in the experiments. Experiment 2—where alternative labels are randomised such that classes have the wrong alternative labels—shows no change compared to the ground truth on any of the syntactic metrics, since changing labels does not influence the structure or richness of the knowledge graph. These results confirm hypothesis 1.

For the Semantic Validation, the Semantic Consistency (SeC) and the Semantic F1 (SeF) scores are lower for experiment 1 compared to the ground truth, whereas the Semantic Interpretability (SeI) is higher. The increase in SeI score can be attributed to the reduction in the number of terms in the knowledge graph. With fewer terms, those with entries in an outside source have a greater impact on the scores. For instance, in this example graph, the occupations *Data Scientist*, *Music Teacher*, and *Data Analyst* have Wikipedia entries, whereas none of the attributes do. In Experiment 1, alternative labels (part of the attributes) are removed, and as these do not have Wikipedia entries, the SeI score increases. However, it is important to note that this score could decrease if attributes with entries are removed, as can be observed in results from other subgraphs. A similar reasoning can be applied to the Semantic F1 score, which relies on keywords from an outside source.

In experiment 2, the Semantic Consistency score lowers compared to the ground truth, which is according to hypothesis 2. The Semantic Consistency also affects the overall quality score, which is slightly lower than the ground truth.

[1]The subgraphs described in Section 4 are available at https://gitlab.com/knowledge-graphs/evaluation. The implementation of the metrics and the keywords can be requested from the authors.

Table 6
Validation results on the pizza subgraph

Exp.	Syntactic Validation			Semantic Validation			
	SyQ	SyR	SyS	SeQ	SeC	SeF	SeI
GT	**0.694**	0.788	0.6	**0.626**	0.965	0.286	0.628
Exp.1	0.549	0.498	0.6	0.599	0.559	0.439	0.8
Exp.2	**0.694**	0.788	0.6	0.623	0.956	0.286	0.628

Table 7
Validation results on the data scientist subgraph

Exp.	Syntactic Validation			Semantic Validation			
	SyQ	SyR	SyS	SeQ	SeC	SeF	SeI
GT	**0.737**	0.974	0.5	0.395	0.999	0.028	0.157
Exp.1	0.713	0.926	0.5	**0.409**	0.986	0.032	0.208
Exp.2	**0.737**	0.974	0.5	0.395	0.999	0.028	0.157

5.2. Pizza subgraph

The results of the validations on the pizza subgraph can be found in Table 6. The results show a similar trend to the example graph, with the Syntactic Quality being lower in experiment 1 due to the Syntactic Richness, and the Semantic Quality being lower in experiment 2 than the Ground Truth due to the Semantic Consistency.

5.3. ESCO subgraphs

5.3.1. ESCO data scientist subgraph

The results of the Syntactic and Semantic Validation for the Data Scientist subgraph can be found in Table 7. The syntactic scores show the same trend as for the previous subgraphs, with the score lowering in experiment 1. For the Semantic Validation, experiment 1 shows the highest scores. This can be explained by the same reason these scores were lower for the example graph: with less terms in the ontology, an (in)consistent group or terms with external sources will have more effect on the scores. There is no difference between the ground truth and experiment 2. The Semantic Consistency score is the same, due to there being only one class, so there are no attribute groups to compare and the resulting score will be the same average from the one attribute group.

5.3.2. ESCO data scientist and data analyst subgraph

The results of the Syntactic and Semantic Validation for the Data Scientist and Data Analyst subgraph can be found in Table 8. The Syntactic scores show a similar pattern as for the previous subgraphs. For the Semantic Validation, the Semantic Interpretation score is highest for experiment 1, leading to the highest Semantic Quality score for this experiment. The Semantic Consistency lowers for both experiment 1 and 2 compared to the ground truth, albeit a small bit. This confirms hypothesis 2. The small differences for this subgraph can be explained by the high similarity in the subgraph; data scientist and data analyst and their attributes are very similar occupations with very similar skill sets.

5.3.3. ESCO systems analysts subgraph

The results of the validations for the ESCO Systems Analysts Subgraph, which is a lot larger than the other subgraphs, can be found in Table 9. The Syntactic Quality shows the same pattern as in the previous subgraphs, confirming hypothesis 1 again. However, the Semantic Validation metrics show no difference between the ground truth and experiment 2. On further inspection of the results, the Semantic Consistency score of experiment 2 is slightly lower than the ground truth: The full score of the Semantic Consistency of the ground truth is 0.9993503996109245, whereas the score on experiment 2 is 0.999265370138205.

Table 8
Validation results on the ESCO data scientist and data analyst subgraph

Exp.	Syntactic Validation			Semantic Validation			
	SyQ	SyR	SyS	SeQ	SeC	SeF	SeI
GT	**0.814**	0.962	0.67	0.394	0.999	0.030	0.151
Exp.1	0.773	0.879	0.67	**0.407**	0.986	0.024	0.211
Exp.2	**0.814**	0.962	0.67	0.393	0.998	0.030	0.151

Table 9
Validation results on the systems analyst subgraph

Exp.	Syntactic Validation			Semantic Validation			
	SyQ	SyR	SyS	SeQ	SeC	SeF	SeI
GT	**0.944**	0.936	0.952	**0.402**	0.999	0.009	0.198
Exp.1	0.849	0.746	0.952	0.394	0.987	0.014	0.18
Exp.2	**0.944**	0.936	0.952	**0.402**	0.999	0.009	0.198

Table 10
Validation results on the diabetes subgraph

Exp.	Syntactic Validation			Semantic Validation			
	SyQ	SyR	SyS	SeQ	SeC	SeF	SeI
GT	**0.643**	0.911	0.38	0.534	0.979	0.033	0.589
Exp.1	0.487	0.598	0.38	**0.599**	0.978	0.035	0.786
Exp.2	**0.643**	0.911	0.38	0.536	0.986	0.033	0.589

Table 11
Validation results on the endocrine subgraph

Exp.	Syntactic Validation			Semantic Validation			
	SyQ	SyR	SyS	SeQ	SeC	SeF	SeI
GT	**0.606**	0.879	0.33	0.557	0.980	0.026	0.667
Exp.1	0.426	0.518	0.33	**0.599**	0.979	0.028	0.792
Exp.2	**0.606**	0.879	0.33	0.559	0.987	0.026	0.667

5.4. MeSH subgraphs

5.4.1. MeSH diabetes subgraph

The results of the validation on the MeSH Diabetes subgraph are shown in Table 10. A similar pattern as before can be observed for the Syntactic Validation, however, for the Semantic Validation, the Semantic Consistency score is higher for experiment 2 than for the ground truth.

5.4.2. MeSH endocrine subgraph

The results of the validations on the MeSH Endocrine Subgraph can be found in Table 11. A similar pattern can be observed as with the Diabetes Subgraph. This might be an indication that for specialised knowledge graphs with complex terms, the Semantic Validation metrics should be adjusted. This will be discussed further in Section 6.

6. Discussion

This work adapted and introduced new evaluation metrics to measure the Syntactic and Semantic Quality of a knowledge graph. The metrics were validated by conducting two experiments, one where labels were removed from a knowledge graph, and a second where alternative labels were randomised. Two hypotheses were proposed: 1. removing information decreases the Syntactic Quality, and 2. randomising labels decreases the Semantic Quality. The results show that the first hypothesis holds for all tested (parts of) knowledge graphs. Removing labels reduced the Syntactic Richness, the metric which measures the average number of attributes and the ratio of non-inheritance relations to all relations. Removing labels, which is removing attributes, logically leads to a lower score on this aspect. The syntactic lawfulness metric was not included due to the selected knowledge graphs not having rules. The Syntactic Structure metric, which calculates the ratio of subclasses to the total number of classes, was also not included in the experiments. However, it can be argued that a similar experiment can be done where subclasses were removed, leading to a similar pattern as the current one where attributes are removed.

The second hypothesis—randomising labels leads to a decrease in the Semantic Quality—did not hold for all subgraphs. The results showed that for the smaller subgraphs, the hypothesis holds, although the difference in score gets smaller as the subgraph gets larger. This becomes clear in the ESCO ontology: the example graph shows a clear difference in Semantic Quality scores, where the score goes down for experiment 2, and the difference becomes smaller as the subgraphs, which are parts of the full knowledge graph or ontology, increase in size. As the graph increases in size, concepts that do not fit in an attribute group have less effect on the Semantic Consistency. Therefore, a power transformation was applied on the results. A larger transformation might be needed for larger ontologies and vice versa. These transformations will need to be tailored to the specific domain of the graph: graphs with greater diversity in concepts will exhibit more variation in scores.

The MeSH taxonomy showed an opposite pattern, where the Semantic Quality actually increased for experiment 2, due to the Semantic Consistency score increasing. This can be explained by looking more closely at the subgraphs. The diabetic subgraph contains classes such as *Diabetic Neuropathies*, and *Diabetic Foot.* Diabetic neuropathy is nerve damage caused by diabetes, and a diabetic foot indicates foot problems caused by diabetes. Examples of alternative labels for these classes in the ground truth are *Diabetic Neuralgia* and *Foot Ulcer, Diabetic*. When randomising the alternative labels for experiment 2, a new alternative label for *Diabetic Neuropathies* is *Diabetic Glomerulosclerosis*: a condition related to kidney function. For *Diabetic Foot*, a randomised alternative label is *Diabetes, Autoimmune* (Diabetes type 1). It can be observed from these labels that all of them contain a form of

the word ‘diabetes’, making them seem similar even though the medical conditions are not alike. When calculating the Semantic Consistency, the similarities between concepts is calculated by using word embeddings, in this case utilising the BERT model [35]. Since BERT is trained on a large corpus of text, it learns to represent words based on their context. In this case, the model does not have medical context besides the phrase, therefore, it ends up giving high similarity scores to terms that have similar word structures, even though they represent different medical conditions. This phenomenon can lead to higher Semantic Consistency scores for experiment 2 compared to the ground truth, despite the incorrectness of the labels.

As the results from the MeSH taxonomy make clear, the Semantic Quality of a knowledge graph relies not only on internal information but also on external sources such as language models and databases like Wikipedia. Thus, scores are influenced by these external sources. While hypothesis 2 was not confirmed by all subgraphs, the metrics remain valid for most cases. A highly specialised knowledge graph such as the MeSH taxonomy might benefit from a language model trained on the medical domain for calculating the similarity within the Semantic Consistency metric (e.g. Med-BERT [101]).

This work did not include experiments to validate the Semantic F1 and Semantic Interpretability, since these metrics are even more dependent on their external source. For the Semantic F1 metric, keywords should be included by a domain expert, or can be extracted from relevant documents using keyword extraction techniques. If there is uncertainty about the quality of the keywords, the weight for this metric could be decreased such that it influences the Semantic Quality score less. The Semantic Interpretability is dependent on its affirmation by external sources. In the implementation of this work, Wikipedia was used, but for ESCO, a labour market related source might be more suitable. Another aspect to consider is similarity in words. Currently, this implementation uses a literal match for keywords and search for outside sources. Ideally, synonyms of the concepts or variations should also be taken into account. For instance, if the plural *Data Scientist**S*** would have been the name of the concept in the example graph, there would not have been a match between the keyword *data scientist* or its Wikipedia page.

While this work focused on evaluating the Syntactic and Semantic Quality of a knowledge graph using existing and newly proposed metrics, there are other possible evaluation metrics that were not included in the experiments. Metrics such as density, which measures the compactness of a knowledge graph, or the size of a knowledge graph could provide additional insights into its quality [54]. Other metrics that can be included from previous work are Pragmatic Quality, which measures the usefulness of the graph, and Social Quality, which measures how well the graph is accepted within a community [81]. Such metrics are dependent on manual evaluation, user feedback, or in some cases, a library of other models. This makes it harder to implement automatic evaluation metrics, and therefore less suitable for change evaluation. An application which includes the graph might support such evaluation metrics. This falls under extrinsic evaluation, whereas the adapted and proposed metrics from this work are intrinsic evaluation [129]. Although such approaches were not part of the experimentation, they could offer valuable insights into knowledge graph quality when coupled with appropriate external sources and manual validation.

Finally, in the experiments, concepts were removed and existing ones shuffled such that they would be incorrect. However, new concepts were not added. This decision was made because evaluating the scores of added concepts would require domain experts. For example, even if the Semantic Quality score improves, it cannot be concluded that this is a positive change without verifying that the added concepts are correct. For future work, such an additional experiment would provide a valuable addition to the applicability of this work.

7. Conclusion

In a rapidly evolving information landscape, the relevance and functionality of knowledge graphs and ontologies rely on their ability to adapt and stay current. This necessity becomes clear in domains such as the labour market and healthcare, where information evolves with innovation. However, manual updating and evaluation of existing knowledge models are labour-intensive tasks, and prone to errors. This study addressed the challenge of evaluating changes in existing knowledge models. Previous research predominantly focused on static models, which resulted in a rich set of evaluation methods. However, not all of them are suitable for changing knowledge graphs, and many of them still require manual work. In this work, evaluation metrics were proposed that are tailored for change evaluation. The metrics were validated through experiments on knowledge graphs from different domains.

This work introduced a comprehensive set of syntactic and semantic metrics designed to evaluate both the structure and content of knowledge graphs. Existing syntactic metrics were expanded to facilitate practical implementation and calculation into a weighted syntactic metric. Additionally, semantic metrics were introduced designed to enable automatic change evaluation.

The metrics were validated with two experiments on several subgraphs: first, by removing concepts from an knowledge graph, and second, by adding faulty ones before measuring the quality using the metrics. The hypotheses suggested that such changes would decrease scores: removing concepts would influence syntactic qualities such as the structure of the model, while adding faulty concepts would affect the consistency of the model, a semantic quality. The results confirm the first hypothesis, indicating that removing concepts mainly decreases the Syntactic Quality of the knowledge model. The results also confirmed the second hypothesis for most subgraphs, where adding faulty concepts decreases the semantic quality. However, for the specialised subgraphs from the medical domain (MeSH), this was not the case. As discussed in the previous section, the semantic quality metrics are dependent on their external source. In the experiments, generic sources were chosen such as the language model BERT and Wikipedia. For a highly specialised knowledge graph such as MeSH, a dedicated language model and reference body might be better suitable.

This work demonstrated the effectiveness of the proposed metrics in evaluating knowledge graphs with dynamic information, particularly when tailored to accommodate variations in knowledge graph size and specialised domains. For future research, exploring the potential of external tools for automatic evaluation would be intriguing. Large Language Models, for example, could play an interesting role in this regard. While the current metrics are best suited for assessing changes in knowledge graphs, as demonstrated in the experiments where scores were compared before and after changes, they can also be applied to static graphs or initial versions. However, interpreting scores in these cases becomes more challenging, since there are no other versions to compare them to. Thus, an investigation into scores across various domains and the fine-tuning of weights would be a valuable addition to this work. Finally, testing an application that utilises a knowledge graph would offer valuable insights into its practical effectiveness and relevance.

Acknowledgements

We gratefully acknowledge the financial support of the TNO ERP program FATE and the supervision of Stephan Raaijmakers. Special thanks go to Quirine Smit, David de Best, and Timo Kooreman, for their valuable input on the implementation of the metrics. Finally, we thank René Bakker for his review of the mathematical formulas and the first draft of this paper.

References

[1] R. Alfred and et al., Ontology-based query expansion for supporting information retrieval in agriculture, in: *The 8th International Conference on Knowledge Management in Organizations*, Springer, 2014, pp. 299–311.

[2] M. Allen, G. Waugh, M. Shaw, S. Tsacoumis, D. Rivkin, P. Lewis, M. Brendle, D. Craven, C. Gregory and D. Connell, The Development and Evaluation of a New O*NET Related Occupations Matrix, Vol. I: Report, *National Center for O* NET Development* (2012).

[3] American Diabetes Association, ABOUT DIABETES: Common Terms, Accessed: May 8, 2024.

[4] G. Angeli, M.J.J. Premkumar and C.D. Manning, Leveraging linguistic structure for open domain information extraction, in: *Proc. of 53 ACL and 7th Int. Joint Conf. on NLP (Vol 1: Long Papers)*, Vol. 1, 2015, pp. 344–354.

[5] M.N. Asim, M. Wasim, M.U.G. Khan, W. Mahmood and H.M. Abbasi, A survey of ontology learning techniques and applications, *Database* **2018** (2018), bay101.

[6] N. Aussenac-Gilles, S. Despres and S. Szulman, The TERMINAE Method and Platform for Ontology Engineering from Texts., 2008.

[7] R.R. Bakker, Knowledge Graphs: Representation and Structuring of Scientific Knowledge, PhD thesis, University of Twente, Enschede, The Netherlands, 1987. ISBN 90-9001963-4.

[8] R. Bakker, R. van Drie, C. Bouter, S. van Leeuwen, L. van Rooijen and J. Top, The Common Greenhouse Ontology: an ontology describing components, properties, and measurements inside the greenhouse, *Engineering Proceedings* **9**(1) (2021), 27.

[9] V.R. Basili and D.M. Weiss, A methodology for collecting valid software engineering data, *IEEE Transactions on software engineering* (1984), 728–738.

[10] R. Bendaoud, A.M.R. Hacene, Y. Toussaint, B. Delecroix and A. Napoli, Text-based ontology construction using relational concept analysis, in: *International Workshop on Ontology Dynamics-IWOD 2007*, 2007.

[11] J.A. Bernabé-Díaz, M. Franco-Nicolas, J.M. Vivo-Molina, M. Quesada-Martínez, A. Duque-Ramos and J.T. Fernandez-Breis, An Automated Process for the Repository-Based Analysis of Ontology Structural Metrics, *IEEE Access* **8** (2020), 148722–148743.

[12] B. Biebow, S. Szulman and A.J. Clément, TERMINAE: A linguistics-based tool for the building of a domain ontology, in: *Int. Conf. on Knowledge Engineering and Knowledge Management*, Springer, 1999, pp. 49–66.

[13] J. Brank, M. Grobelnik and D. Mladenic, A survey of ontology evaluation techniques, in: *Proceedings of the conference on data mining and data warehouses (SiKDD 2005)*, Citeseer, 2005, pp. 166–170.

[14] C.A. Brewster, Mind the gap: Bridging from text to ontological knowledge, PhD thesis, University of Sheffield, 2008.

[15] C. Brewster, H. Alani, S. Dasmahapatra and Y. Wilks, Data driven ontology evaluation, *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)* (2004).

[16] P. Buitelaar, D. Olejnik and M. Sintek, A protégé plug-in for ontology extraction from text based on linguistic analysis, in: *European Semantic Web Symposium*, Springer, 2004, pp. 31–44.

[17] A. Burton-Jones, V.C. Storey, V. Sugumaran and P. Ahluwalia, A semiotic metrics suite for assessing the quality of ontologies, *Data & Knowledge Engineering* **55**(1) (2005), 84–102.

[18] M. Cheatham and P. Hitzler, String similarity metrics for ontology alignment, in: *International semantic web conference*, Springer, 2013, pp. 294–309.

[19] J. Chen, E. Jiménez-Ruiz, I. Horrocks, D. Antonyrajah, A. Hadian and J. Lee, Augmenting ontology alignment by semantic embedding and distant supervision, in: *European Semantic Web Conference*, Springer, 2021, pp. 392–408.

[20] M. Cifuentes, J. Boyer, D.A. Lombardi and L. Punnett, Use of O* NET as a job exposure matrix: a literature review, *American journal of industrial medicine* **53**(9) (2010), 898–914.

[21] P. Cimiano and J. Völker, Text2onto: A framework for ontology learning and data-driven change discovery, in: *International conference on application of natural language to information systems*, Springer, 2005, pp. 227–238.

[22] P. Cimiano, A. Mädche, S. Staab and J. Völker, Ontology learning, in: *Handbook on ontologies*, Springer, 2009, pp. 245–267.

[23] L. Cui, F. Wei and M. Zhou, Neural Open Information Extraction, *arXiv preprint arXiv:1805.04270* (2018).

[24] H. Cunningham, D. Maynard, K. Bontcheva and V. Tablan, GATE: an architecture for development of robust HLT applications, in: *Proceedings of the 40th annual meeting on association for computational linguistics*, Association for Computational Linguistics, 2002, pp. 168–175.

[25] M. de Boer, K. Schutte and W. Kraaij, Knowledge based query expansion in complex multimedia event detection, *Multimedia Tools and Applications* **75**(15) (2016), 9025–9043.

[26] M.H. de Boer and et al., Query Interpretation–an application of semiotics in image retrieval, *International Journal on Advances in Software* **3 & 4** (2015), 435–449.

[27] M.H. De Boer, Y.-J. Lu, H. Zhang, K. Schutte, C.-W. Ngo and W. Kraaij, Semantic reasoning in zero example video event retrieval, *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* **13**(4) (2017), 60.

[28] M. de Boer and J. Verhoosel, Creating Data-Driven Ontologies: An Agriculture Use Case, in: *ALLDATA 2019: the Fifth International Conference on Big Data, Small Data, Linked Data and Open Data, Valencia, Spain 24-28 March 2019, 52-57*, 2019.

[29] M.-C. De Marneffe, B. MacCartney and C.D. Manning, Generating typed dependency parses from phrase structure parses (2006).

[30] J. De Smedt, M. le Vrang and A. Papantoniou, ESCO: Towards a Semantic Web for the European Labor Market, in: *Ldow@ www*, 2015.

[31] T. Declerck, A set of tools for integrating linguistic and non-linguistic information, in: *Proceedings of SAAKM (ECAI Workshop)*, 2002.

[32] A. Degbelo, A snapshot of ontology evaluation criteria and strategies, in: *Proceedings of the 13th International Conference on Semantic Systems*, 2017, pp. 1–8.

[33] L. Del Corro and R. Gemulla, Clausie: clause-based open information extraction, in: *Proceedings of the 22nd international conference on World Wide Web*, ACM, 2013, pp. 355–366.

[34] L. Den Boer, Normalized ArcTan S-Function - a versatile smoothing function, 2011. doi:10.13140/RG.2.2.32580.94087.

[35] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *arXiv preprint arXiv:1810.04805* (2018).

[36] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. doi:10.18653/v1/N19-1423.

[37] R. Dividino, M. Romanelli and D. Sonntag, Semiotic-based Ontology Evaluation Tool (S-OntoEval)., in: *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, 2008.

[38] D. Dou, H. Wang and H. Liu, Semantic data mining: A survey of ontology-based approaches, in: *Semantic Computing (ICSC), 2015 IEEE International Conference on*, IEEE, 2015, pp. 244–251.

[39] A. Duque-Ramos, J.T. Fernández-Breis, M. Iniesta, M. Dumontier, M.E. Aranguren, S. Schulz, N. Aussenac-Gilles and R. Stevens, Evaluation of the OQuaRE framework for ontology quality, *Expert Systems with Applications* **40**(7) (2013), 2696–2703.

[40] L. Ehrlinger and W. Wöß, Towards a definition of knowledge graphs., *SEMANTiCS (Posters, Demos, SuCCESS)* **48**(1–4) (2016), 2.

[41] A. Fader, S. Soderland and O. Etzioni, Identifying relations for open information extraction, in: *Proceedings of the conference on empirical methods in natural language processing*, Association for Computational Linguistics, 2011, pp. 1535–1545.

[42] D. Faria, C. Pesquita, E. Santos, M. Palmonari, I.F. Cruz and F.M. Couto, The AgreementMakerLight Ontology Matching System, in: *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*, Springer, 2013, pp. 527–541.

[43] M. Faruqui, J. Dodge, S.K. Jauhar, C. Dyer, E. Hovy and N.A. Smith, Retrofitting word vectors to semantic lexicons, *arXiv preprint arXiv:1411.4166* (2014).

[44] R. FB, Medical subject headings., *Bulletin of the Medical Library Association* **51** (1963), 114–116.

[45] E.A. Feigenbaum, The art of artificial intelligence: themes and case studies of knowledge engineering, in: *Proceedings of the 5th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'77, Morgan Kaufmann Publishers Inc., 1977, pp. 1014–1029–.

[46] E.A. Feigenbaum, Knowledge Engineering, *Annals of the New York Academy of Sciences* **426**(1) (1984), 91–107.

[47] M. Franco, J.M. Vivo, M. Quesada-Martínez, A. Duque-Ramos and J.T. Fernández-Breis, Evaluation of ontology structural metrics based on public repository data, *Briefings in bioinformatics* **21**(2) (2020), 473–485.

[48] A. Gangemi and V. Presutti, Ontology design patterns, in: *Handbook on ontologies*, Springer, 2009, pp. 221–243.

[49] A. Gangemi, V. Presutti, D. Reforgiato Recupero, A.G. Nuzzolese, F. Draicchio and M. Mongiovì, Semantic web machine reading with FRED, *Semantic Web* **8**(6) (2017), 873–893.

[50] S. Gillani and A. Kő, ProMine: a text mining solution for concept extraction and filtering, in: *Corporate Knowledge Discovery and Organizational Learning*, Springer, 2016, pp. 59–82.

[51] R. Glauber and D.B. Claro, A systematic mapping study on open information extraction, *Expert Systems with Applications* **112** (2018), 372–387.

[52] M. Grootendorst, Keyword Extraction with BERT, 2020, Accessed on: 23-08-2023. https://towardsdatascience.com/keyword-extraction-with-bert-724efca412ea.

[53] N. Guarino, D. Oberle and S. Staab, What is an ontology?, in: *Handbook on ontologies*, Springer, 2009, pp. 1–17.

[54] C. Guéret, P. Groth, C. Stadler and J. Lehmann, Assessing linked data mappings using network measures, in: *The Semantic Web: Research and Applications: 9th Extended Semantic Web Conference, ESWC 2012, Heraklion, Crete, Greece, May 27-31, 2012. Proceedings 9*, Springer, 2012, pp. 87–102.

[55] A. Gyrard, G. Atemezing and M. Serrano, PerfectO: an online toolkit for improving quality, accessibility, and classification of domain-based ontologies, in: *Semantic IoT: Theory and Applications*, Springer, 2021, pp. 161–192.

[56] Y. He, J. Chen, D. Antonyrajah and I. Horrocks, BERTMap: A BERT-based Ontology Alignment System, *arXiv preprint arXiv:2112.02682* (2021).

[57] M.A. Hearst, Automatic acquisition of hyponyms from large text corpora, in: *Proceedings of the 14th conference on Computational linguistics-Volume 2*, Association for Computational Linguistics, 1992, pp. 539–545.

[58] N. Heist, S. Hertling and H. Paulheim, KGrEaT: a framework to evaluate knowledge graphs via downstream tasks, in: *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 2023, pp. 3938–3942.

[59] H. Hlomani and D. Stacey, Approaches, methods, metrics, measures, and subjectivity in ontology evaluation: A survey, *Semantic Web Journal* **1**(5) (2014), 1–11.

[60] H. Hlomani and D.A. Stacey, Contributing evidence to data-driven ontology evaluation workflow ontologies perspective, in: *5th International Conference on Knowledge Engineering and Ontology Development, KEOD 2013*, 2013, pp. 207–213.

[61] A. Hogan, E. Blomqvist, M. Cochez, C. d'Amato, G.D. Melo, C. Gutierrez, S. Kirrane, J.E.L. Gayo, R. Navigli, S. Neumaier et al., Knowledge graphs, *ACM Computing Surveys (Csur)* **54**(4) (2021), 1–37.

[62] V. Iyer, A. Agarwal and H. Kumar, VeeAlign: a supervised deep learning approach to ontology alignment., in: *OM@ ISWC*, 2020, pp. 216–224.

[63] V. Iyer, L.M. Sanagavarapu and Y. Raghu Reddy, A framework for syntactic and semantic quality evaluation of ontologies, in: *International Conference On Secure Knowledge Management In Artificial Intelligence Era*, Springer, 2021, pp. 73–93.

[64] X. Jiang and A.-H. Tan, CRCTOL: A semantic-based domain ontology learning system, *Journal of the American Society for Information Science and Technology* **61**(1) (2010), 150–168.

[65] E. Jiménez-Ruiz and B. Cuenca Grau, Logmap: Logic-based and Scalable Ontology Matching, in: *International Semantic Web Conference*, Springer, 2011, pp. 273–288.

[66] A. Kahlawi, An Ontology Driven ESCO LOD Quality Enhancement, *International Journal of Advanced Computer Science and Applications (IJACSA)* **11**(3) (2020).

[67] Y.-B. Kang, P.D. Haghighi and F. Burstein, CFinder: An intelligent key concept finder from text for ontology development, *Expert Systems with Applications* **41**(9) (2014), 4494–4504.

[68] M. Klein and N.F. Noy, A component-based framework for ontology evolution, in: *Workshop on Ontologies and Distributed Systems at IJCAI*, Vol. 3, 2003, p. 4.

[69] M.C. Klein and D. Fensel, Ontology versioning on the Semantic Web., in: *SWWS*, 2001, pp. 75–91.

[70] P. Kolyvakis, A. Kalousis and D. Kiritsis, Deepalignment: Unsupervised ontology matching with refined word vectors, in: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2018, pp. 787–798.

[71] H. Kondylakis, A. Nikolaos, P. Dimitra, K. Anastasios, K. Emmanouel, K. Kyriakos, S. Iraklis, K. Stylianos and N. Papadakis, Delta: A Modular Ontology Evaluation System, *Information* **12**(8) (2021), 301.

[72] K.I. Kotis, G.A. Vouros and D. Spiliotopoulos, Ontology engineering methodologies for the evolution of living and reused ontologies: status, trends, findings and recommendations, *The Knowledge Engineering Review* **35** (2020), e4.

[73] S. Kumar, A survey of deep learning methods for relation extraction, *arXiv preprint arXiv:1705.03645* (2017).

[74] Y. Lin, S. Shen, Z. Liu, H. Luan and M. Sun, Neural relation extraction with selective attention over instances, in: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1, 2016, pp. 2124–2133.

[75] C.E. Lipscomb, Medical subject headings (MeSH), *Bulletin of the Medical Library Association* **88**(3) (2000), 265.

[76] R. Lourdusamy and A. John, A review on metrics for ontology evaluation, in: *2018 2nd International Conference on Inventive Systems and Control (ICISC)*, IEEE, 2018, pp. 1415–1421.

[77] A. Lozano-Tello and A. Gómez-Pérez, Ontometric: A method to choose the appropriate ontology, *Journal of Database Management (JDM)* **15**(2) (2004), 1–18.

[78] M. Mao, Y. Peng and M. Spring, Ontology mapping: as a binary classification problem, *Concurrency and Computation: Practice and Experience* **23**(9) (2011), 1010–1025.

[79] Y. Matsuo and M. Ishizuka, Keyword extraction from a single document using word co-occurrence statistical information, *International Journal on Artificial Intelligence Tools* **13**(01) (2004), 157–169.

[80] M. McDaniel and V.C. Storey, Evaluating domain ontologies: clarification, classification, and challenges, *ACM Computing Surveys (CSUR)* **52**(4) (2019), 1–44.

[81] M. McDaniel, V.C. Storey and V. Sugumaran, Assessing the quality of domain ontologies: Metrics and an automated ranking system, *Data & Knowledge Engineering* **115** (2018), 32–47.

[82] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado and J. Dean, Distributed representations of words and phrases and their compositionality, in: *Adv. in neural information processing systems*, 2013, pp. 3111–3119.

[83] G.A. Miller, WordNet: a lexical database for English, *Communications of the ACM* **38**(11) (1995), 39–41.

[84] G.A. Miller, WordNet: a lexical database for English, *Communications of the ACM* **38**(11) (1995), 39–41.

[85] S. Mittal, A. Joshi, T. Finin et al., Thinking, Fast and Slow: Combining Vector Spaces and Knowledge Graphs, *arXiv* (2017).

[86] N. Mrkšić, D.O. Séaghdha, B. Thomson, M. Gašić, L. Rojas-Barahona, P.-H. Su, D. Vandyke, T.-H. Wen and S. Young, Counter-fitting word vectors to linguistic constraints, *arXiv preprint arXiv:1603.00892* (2016).

[87] S.J. Nelson, W.D. Johnston and B.L. Humphreys, Relationships in medical subject headings (MeSH), in: *Relationships in the Organization of Knowledge*, Springer, 2001, pp. 171–184.

[88] D. Ngo and Z. Bellahsene, Overview of YAM++—(not) Yet Another Matcher for ontology alignment task, *Journal of Web Semantics* **41** (2016), 30–49.

[89] C. Niklaus, M. Cetto, A. Freitas and S. Handschuh, A Survey on Open Information Extraction, *arXiv preprint arXiv:1806.05599* (2018).

[90] I. Nkisi-Orji, N. Wiratunga, S. Massie, K.-Y. Hui and R. Heaven, Ontology alignment based on word embedding and random forest classification, in: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, Springer, 2018, pp. 557–572.

[91] P. Øhrstrøm, J. Andersen and H. Schärfe, What has happened to ontology, in: *International Conference on Conceptual Structures*, Springer, 2005, pp. 425–438.

[92] L. Ouyang, B. Zou, M. Qu and C. Zhang, A method of ontology evaluation based on coverage, cohesion and coupling, in: *Fuzzy Systems and Knowledge Discovery (FSKD), 2011 Eighth International Conference on*, Vol. 4, IEEE, 2011, pp. 2451–2455.

[93] J. Park, W. Cho and S. Rho, Evaluating ontology extraction tools using a comprehensive evaluation framework, *Data & Knowledge Engineering* **69**(10) (2010), 1043–1061.

[94] İ. Pembeci, Using word embeddings for ontology enrichment, *International Journal of Intelligent Systems and Applications in Engineering* **4**(3) (2016), 49–56.

[95] P. Peña, R.T. Lado, R. Del Hoyo, M. del Carmen Rodríguez-Hernández and D. Abadía-Gallego, Ontology-quality Evaluation Methodology for Enhancing Semantic Searches and Recommendations: A Case Study., in: *WEBIST*, 2020, pp. 277–284.

[96] N.G. Peterson, M.D. Mumford, W.C. Borman, P. Jeanneret and E.A. Fleishman, *An occupational information system for the 21st century: The development of O* NET.*, American Psychological Association, 1999.

[97] E. Pianta and S. Tonelli, KX: A flexible system for keyphrase extraction, in: *Proceedings of the 5th international workshop on semantic evaluation*, Association for Computational Linguistics, 2010, pp. 170–173.

[98] H. Poon and P. Domingos, Unsupervised ontology induction from text, in: *Proceedings of the 48th annual meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, 2010, pp. 296–305.

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

[100] L. Qiu, J. Yu, Q. Pu and C. Xiang, Knowledge entity learning and representation for ontology matching based on deep neural networks, *Cluster Computing* **20**(2) (2017), 969–977.

[101] L. Rasmy, Y. Xiang, Z. Xie, C. Tao and D. Zhi, Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction, *NPJ digital medicine* **4**(1) (2021), 86.

[102] A. Rector, N. Drummond, M. Horridge, J. Rogers, H. Knublauch, R. Stevens, H. Wang and C. Wroe, Pizza.owl, Accessed: 2023-08-25.

[103] A. Rector, N. Drummond, M. Horridge, J. Rogers, H. Knublauch, R. Stevens, H. Wang and C. Wroe, OWL Pizzas: Common errors & common patterns from practical experience of teaching OWL-DL, in: *Proceedings of the European Knowledge Acquisition Workshop (EKAW-2004)*, Springer Verlag, Northampton, England, 2004.

[104] A. Reiz and K. Sandkuhl, Design Decisions and Their Implications: An Ontology Quality Perspective, in: *Perspectives in Business Informatics Research: 19th International Conference on Business Informatics Research, BIR 2020, Vienna, Austria, September 21–23, 2020, Proceedings 19*, Springer, 2020, pp. 111–127.

[105] A. Reiz and K. Sandkuhl, NEOntometrics: A Flexible and Scalable Software for Calculating Ontology Metrics, in: *18th International Conference on Semantic Systems, SEMPDW 2022, 13 September 2022 through 15 September 2022*, CEUR-WS, 2022.

[106] A. Reiz, H. Dibowski, K. Sandkuhl and B. Lantow, Ontology Metrics as a Service (OMaaS)., in: *KEOD*, 2020, pp. 250–257.

[107] ResumeWorded, Data Scientist Resume Keywords and Skills (Hard Skills), Accessed: May 8, 2024.

[108] M. Rospocher, S. Tonelli, L. Serafini and E. Pianta, Corpus-based terminological evaluation of ontologies, *Applied Ontology* **7**(4) (2012), 429–448.
[109] M. Schlichtkrull, T.N. Kipf, P. Bloem, R. van den Berg, I. Titov and M. Welling, Modeling relational data with graph convolutional networks, in: *European Semantic Web Conference*, Springer, 2018, pp. 593–607.
[110] M. Schmitz, R. Bart, S. Soderland, O. Etzioni et al., Open language learning for information extraction, in: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Association for Computational Linguistics, 2012, pp. 523–534.
[111] sci-kit learn team, CountVectorizer, Accessed: 2019-07-25.
[112] A. Shakya and S. Paudel, Job-candidate matching using ESCO ontology, *Journal of the Institute of Engineering* **15**(1) (2019), 1–13.
[113] P. Shvaiko and J. Euzenat, Ontology matching: state of the art and future challenges, *IEEE Transactions on knowledge and data engineering* **25**(1) (2011), 158–176.
[114] M.K. Smith, OWL Web ontology language guide, W3C recommendation, *http://www. w3. org/TR/2004/REC-owl-guide-20040210/* (2004).
[115] M. Song, I.-Y. Song, X. Hu and R.B. Allen, Integration of association rules and ontologies for semantic query expansion, *Data & Knowledge Engineering* **63**(1) (2007), 63–75.
[116] R. Speer and C. Havasi, Representing General Relational Knowledge in ConceptNet 5., in: *LREC*, 2012, pp. 3679–3686.
[117] P. Spyns, EvaLexon: Assessing Triples M ined from Texts, *STAR* **9** (2005), 09.
[118] L. Stojanovic, Methods and tools for ontology evolution, PhD thesis, University of Karlsruhe, 2004.
[119] S. Tartir and I.B. Arpinar, Ontology evaluation and ranking using OntoQA, in: *International conference on semantic computing (ICSC 2007)*, IEEE, 2007, pp. 185–192.
[120] The Endocrine Society, Glossary, Accessed: May 8, 2024.
[121] A. Tibaut and S. Guerra de Oliveira, A Framework for the Evaluation of the Cultural Heritage Information Ontology, *Applied Sciences* **12**(2) (2022), 795.
[122] I. Tiddi, N.B. Mustapha, Y. Vanrompay and M.-A. Aufaure, Ontology learning from open linked data and web snippets, in: *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*, Springer, 2012, pp. 434–443.
[123] S. Tsaneva, Evaluating Knowledge Graphs with Hybrid Intelligence, in: *European Semantic Web Conference*, Springer, 2023, pp. 310–320.
[124] S. Verberne, M. Sappelli, D. Hiemstra and W. Kraaij, Evaluation and analysis of term scoring methods for term extraction, *Information Retrieval Journal* **19**(5) (2016), 510–545.
[125] J. Wales and L. Sanger, Wikipedia, Accessed on 01-05-2024.
[126] L.L. Wang, C. Bhagavatula, M. Neumann, K. Lo, C. Wilhelm and W. Ammar, Ontology alignment in the biomedical domain using entity definitions and context, *arXiv preprint arXiv:1806.07976* (2018).
[127] R. Whately, *Elements of logic*, Longman, Green, Longman, Roberts and Green, 1897.
[128] R.S.I. Wilson, J.S. Goonetillake, A. Ginige and W.A. Indika, Ontology quality evaluation methodology, in: *International Conference on Computational Science and Its Applications*, Springer, 2022, pp. 509–528.
[129] R. Wilson, J.S. Goonetillake, W. Indika and A. Ginige, Analysis of ontology quality dimensions, criteria and metrics, in: *International Conference on Computational Science and Its Applications*, Springer, 2021, pp. 320–337.
[130] R. Wilson, J. Goonetillake, W. Indika and A. Ginige, A conceptual model for ontology quality assessment, *Semantic Web* (2022), 1–47.
[131] S.I. Wilson, J.S. Goonetillake, A. Ginige and A.I. Walisadeera, Towards a usable ontology: the identification of quality characteristics for an ontology-driven decision support system, *IEEE Access* **10** (2022), 12889–12912.
[132] G. Wohlgenannt and F. Minic, Using word2vec to Build a Simple Ontology Learning System, in: *International Semantic Web Conference (Posters & Demos)*, 2016.
[133] F. Wu and D.S. Weld, Open information extraction using Wikipedia, in: *Proceedings of the 48th annual meeting of the association for computational linguistics*, Association for Computational Linguistics, 2010, pp. 118–127.
[134] C. Xiang, T. Jiang, B. Chang and Z. Sui, ERSOM: A structural ontology matching approach using automatically learned entity representation, in: *Proceedings of the 2015 conference on empirical methods in natural language processing*, 2015, pp. 2419–2429.
[135] H. Yao, A.M. Orme and L. Etzkorn, Cohesion metrics for ontology design and application, *Journal of Computer science* **1**(1) (2005), 107–113.
[136] A. Yates and et al., Textrunner: open information extraction on the web, in: *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, Association for Computational Linguistics, 2007, pp. 25–26.
[137] J. Yu, J.A. Thom and A. Tam, Requirements-oriented methodology for evaluating ontologies, *Information Systems* **34**(8) (2009), 766–791.
[138] F. Zablith, Evolva: A comprehensive approach to ontology evolution, in: *European Semantic Web Conference*, Springer, 2009, pp. 944–948.
[139] F. Zablith, G. Antoniou, M. d'Aquin, G. Flouris, H. Kondylakis, E. Motta, D. Plexousakis and M. Sabou, Ontology evolution: a process-centric survey, *The knowledge engineering review* **30**(1) (2015), 45–75.
[140] Y. Zhang, X. Wang, S. Lai, S. He, K. Liu, J. Zhao and X. Lv, Ontology matching with word embeddings, in: *Chinese computational linguistics and natural language processing based on naturally annotated big data*, Springer, 2014, pp. 34–45.
[141] A. Zouaq, An overview of shallow and deep natural language processing for ontology learning, in: *Ontology learning and knowledge discovery using the web: Challenges and recent advances*, IGI Global, 2011, pp. 16–37.