Nigerian Journal of Operations Research (NIJOR)

https://nijor.org.ng/
pISSN: 2141-7938
eISSN: 3156-5956

A Formal Ontology Evaluation Methodology for Education-Domain Knowledge Graphs

Manunyi Godfrey[1], Shaphat Safeton[2,*], Kayode Oladipupo Oluborode[3]

[1,2,3] Computer Science, Computing, Modibbo Adama State, Yola, Nigeria

ARTICLE INFO

Article history:
Received 26 August, 2026
Revised form 23 September, 2026
Accepted 27 September, 2026
Available online 29 September, 2026

Keywords: Ontology Evaluation; Knowledge Graphs; Education Domain; Ontology Quality Metrics; OQuaRE; Bloom's Taxonomy; Semantic Web; Formal Methods

ABSTRACT

Knowledge graphs built on domain ontologies increasingly underpin adaptive learning platforms, learning-analytics dashboards, and personalized recommendation systems in education. The quality of the underlying ontology directly determines the reliability of the resulting knowledge graph and, consequently, the pedagogical decisions derived from it. Yet most existing ontology-evaluation frameworks were designed for general-purpose, biomedical, or enterprise ontologies and do not explicitly account for pedagogical constructs such as alignment with Bloom's/Anderson's taxonomy, learning-objective traceability, or curriculum granularity. This paper proposes a formal, multi-layer methodology for evaluating education-domain ontologies and the knowledge graphs derived from them. The methodology combines structural, logical, functional, pedagogical-fitness, and usability layers into a single weighted quality score, with weights elicited through stakeholder-driven requirements analysis. We review the state of the art in general ontology evaluation and in education-domain ontology development, formalize the proposed scoring model, and illustrate its application through a case study on a syllabus and learning-objective knowledge graph. The methodology gives ontology engineers and instructional designers a reproducible, criterion-based instrument for assessing and improving education-domain ontologies before deployment in production learning systems.

1. Introduction

Across the education sector, ontologies and the knowledge graphs built on top of them are increasingly used to represent syllabuses, learning objectives, assessment items, and the relationships between them, enabling course recommendation, adaptive learning-path generation, and fine-grained learning analytics. A growing body of work has surveyed how ontologies are constructed and used in the e-learning domain (Khoiruddin *et al.*, 2023; Wang & Wang, 2021), documenting a wide diversity of modelling choices, granularities, and intended uses. Because downstream applications (recommendation engines, adaptive sequencing, and dashboards that report a learner's mastery of specific objectives) inherit whatever errors, omissions, or inconsistencies exist in the ontology, the quality of the ontology is not a secondary concern; it is a precondition for trustworthy educational technology.

* *Corresponding author.*
E-mail address: safethon@gmail.com

https://doi.org/10.67868/nijor/0303.026

A substantial ontology-evaluation literature already exists outside the education domain, offering criteria, layers, and tool-supported frameworks such as OQuaRE, ROMEO, Delta, and FOCA. These general frameworks are valuable but are not designed around the specific semantics of instructional content: they do not, by default, check whether learning objectives are traceable to Bloom's-taxonomy levels, whether assessment items are correctly linked to the objectives they measure, or whether the granularity of a curriculum ontology is appropriate for adaptive sequencing. Reviews of e-learning ontology practice repeatedly note that evaluation is the least mature stage of the ontology life cycle in this domain (Heiyanthuduwage, 2022), and that most published education ontologies are validated informally, if at all, before being embedded in production systems (Novogrudska & Popova, 2021). A comparable gap between formal and informal quality control has been documented for research quality assurance in Nigerian public universities (Watsilla, 2026). Recent work on LLM-assisted ontology evaluation likewise describes competency-question verification as costly, labour-intensive, and error-prone (Lippolis *et al.*, 2025b).

This paper addresses that gap by proposing a formal, multi-layer methodology for evaluating education-domain ontologies and the knowledge graphs instantiated from them. The contributions of the paper are fourfold: (1) a synthesis of general-purpose ontology-evaluation criteria, approaches, and layers relevant to the education domain; (2) a review of ontology-evaluation practice specific to e-learning and instructional-design ontologies; (3) a formal scoring model in which every layer (structural, logical, functional, pedagogical-fitness, and usability) has its own explicit, normalized mathematical definition (Eqs. (2)–(13) and the layers are integrated into a single weighted, reproducible ontology-quality score (Eq. (1); and (4) a documented stakeholder weight-elicitation protocol with an explicit aggregation rule and saturation stopping criterion (Section 4.3), demonstrated on a case study whose graph dimensions, competency questions, and SPARQL queries are reported so that the evaluation can be repeated (Section 5 and Appendix A). Because the composite score is a multi-criteria aggregate whose weights come from stakeholders, the approach is close in spirit to the multi-criteria decision-making studies published in this journal (Barma *et al.*, 2026), and the graph descriptors used to characterize the evaluated knowledge graph follow the graph-theoretic modelling vocabulary used in recent NIJOR work (Babarinsa *et al.*, 2026; Ojo & Babarinsa, 2026). The remainder of the paper is organized as follows. Section 2 reviews general ontology-evaluation approaches. Section 3 reviews evaluation practice specific to education-domain ontologies. Section 4 formalizes the proposed methodology, including the layer definitions and the weight-elicitation protocol. Section 5 applies the methodology to a case study. Section 6 concludes, states limitations, and outlines future work. Appendix A lists sample competency questions and their SPARQL formalizations.

2. Literature Review

Ontology evaluation has been approached from several complementary angles, which several surveys organize into criteria, approaches, and layers (Fathian Dastgerdi, 2012). Requirement-oriented methodologies such as ROMEO evaluate an ontology against the competency questions and requirements that motivated its construction (Yu *et al.*, 2009; Grüninger & Fox, 1995). Layered conceptual models decompose ontology quality into distinct dimensions, for example structural, functional, and usability-related quality, and associate each dimension with concrete criteria and metrics (Wilson *et al.*, 2023; Wilson *et al.*, 2021). Tool-supported systems operationalize these criteria: Delta implements a modular pipeline of evaluation modules that can be combined depending on the evaluation goal (Kondylakis *et al.*, 2021), while the OQuaRE framework adapts software-quality standards (ISO/IEC 25000 SQuaRE) to ontologies and has been validated both on ontology-design guidelines (Duque-Ramos *et al.*, 2014) and independently re-evaluated for its metric reliability (Duque-Ramos *et al.*, 2013). Comparative methodologies extend single-ontology evaluation to the problem of comparing several candidate ontologies against the same requirements, as demonstrated

for general-purpose ontologies (Cardinale *et al.*, 2021) and for SLAM ontologies specifically (Cardinale *et al.*, 2020). Recent domain-specific evaluations continue to pair structural base, schema, and graph metrics with requirement-derived criteria such as completeness, coverage, availability, and adaptability (Norouzi *et al.*, 2024).

A second stream of work focuses on extracting and selecting the criteria themselves. Criteria have been derived inductively from the literature and validated empirically (Kim & Oh, 2019), and criteria-selection frameworks have been proposed to help practitioners choose an appropriate subset of criteria for a given evaluation context (Hooi *et al.*, 2015). Full-ontology evaluation systems such as FOEval combine multiple criteria into a single evaluation pipeline (Bouiadjra & Benslimane, 2011a), and the FOCA methodology (Bandeira *et al.*, 2016) is organized around the Goal–Question–Metric (GQM) approach: the evaluator first fixes the type of ontology (domain, task, or application ontology), then iterates through goals, questions, and metrics, and finally obtains a quality estimate from a statistical model. Practically oriented methods aimed at practitioners rather than researchers have also been proposed (López Rodríguez *et al.*, 2018), as have expert-in-the-loop approaches that combine automated metrics with domain-expert judgement (Rodriguez Apodaca *et al.*, 2019).

A distinct line of research emphasizes the social and community dimension of ontology quality: reuse potential, adoption, and the role of the ontology-engineering community in validating an ontology (Talebpour *et al.*, 2020; Talebpour, 2020), including empirical interview studies of how practitioners actually judge ontologies for reuse (Talebpour *et al.*, 2017). Ermolayev and colleagues' OntoElect line of work treats evaluation as an iterative, stakeholder-driven election process, validating scope ontologies against evolving stakeholder requirements (Tatarintseva *et al.*, 2013a), formalizing the requirements-elicitation step itself (Ermolayev, 2018), and quantifying ontology fitness using saturation- and vote-based metrics that track how requirements stabilize over successive iterations (Tatarintseva *et al.*, 2013b). Complementing these methodological contributions, several papers step back to characterize the field as a whole: cataloguing the persistent challenges in ontology evaluation (Gelernter & Jha, 2016), providing a concise snapshot of criteria and strategies in current use (Degbelo, 2017), and tracing the lineage of the field back to early frameworks for verifying knowledge-sharing technology (Gómez-Pérez, 1996).

Metrics themselves have been the subject of dedicated reviews, covering structural and lexical metrics for automated evaluation (Lourdusamy & John, 2018), broader surveys of methods and tools (Lourdusamy & John, 2017), and lightweight approaches designed to check ontology sufficiency without the overhead of a full evaluation cycle (Sanagavarapu *et al.*, 2017). Comparative studies of evaluation-methodology design choices (Silva-López *et al.*, 2020), periodic surveys of research advances in the field (Lei, 2016), and dedicated surveys of ontology metrics (García *et al.*, 2010) together sketch a maturing but fragmented landscape, a fragmentation echoed by recent surveys of currently available evaluation tools (Debnath & Patel, 2023), papers positioning new approaches against the state of the art (Bouiadjra & Benslimane, 2011b), and encyclopaedic treatments of ontological evaluation and validation as a topic in its own right (Tartir *et al.*, 2010). Finally, module-level evaluation metrics address the increasingly common case where ontologies are composed of reusable modules rather than monolithic artefacts (Khan & Keet, 2016); critical analyses of ontology-construction life cycles situate evaluation as one stage among several and highlight how weakly it is integrated with the earlier design stages in practice (Nefzi *et al.*, 2014); clarification-and-classification papers attempt to bring conceptual order to the terminology used across this literature (McDaniel & Storey, 2019); and dedicated studies of ontology fitness against formalized stakeholder requirements (Vasileyko & Ermolayev, 2018) and of graph-independent structural evaluation branches (Amirhosseini & Salim, 2019) round out the general picture. Comparative-metrics work by Hicks further underlines that different metrics can rank the same set of candidate ontologies inconsistently, reinforcing the need for explicit, task-appropriate criteria rather than a single universal score (Hicks,

2017). At the knowledge-graph level, comprehensive surveys now cover quality dimensions and metrics together with assessment, error detection, and correction (Xue & Zou, 2023).

To our knowledge, none of the frameworks reviewed above defines a layer for pedagogical constructs, and none combines layer scores with a documented, auditable stakeholder-weighting protocol. Requirement-oriented and OntoElect-style methods elicit requirements but do not prescribe a numeric aggregation; OQuaRE and Delta supply metrics but not education-specific ones; FOCA derives a quality estimate from a statistical model rather than from stakeholder weights. The novelty of this paper is therefore the combination of (i) an education-specific pedagogical-fitness layer with explicit formulas, (ii) formal per-layer definitions that make the composite score reproducible by a third party, and (iii) an elicitation protocol with a stated aggregation rule and stopping criterion. This choice is consistent with recent evidence that the importance of quality dimensions depends on the task and the people involved, so that equal or assumed weights are poorly justified for composite quality indicators (Mohammed *et al.*, 2025; Pellegrino *et al.*, 2026).

3. Education-Domain Ontologies and Their Evaluation Needs

Within the education domain specifically, ontology-evaluation practice has developed somewhat independently of the general literature reviewed above. Early work characterized the properties that make a domain ontology suitable for web-based learning and proposed quality-evaluation criteria tailored to that context (Tankeleviciene & Damasevicius, 2009). The Knowledge Puzzle project evaluated automatically generated domain ontologies against human-built gold standards, an approach still influential in automated e-learning ontology generation (Zouaq & Nkambou, 2009). More recent recommendation-oriented ontologies, such as EduCOR (Ilkou *et al.*, 2021), explicitly integrate career- and competency-oriented concepts alongside curriculum concepts, illustrating the trend toward richer, application-driven education ontologies that a generic structural metric alone cannot adequately assess. A recent curriculum-ontology study extracts courses, topics, learning objectives, and cognitive levels from curriculum documents and validates the result with retrieval precision and recall together with expert ratings of accuracy, coverage, and sufficiency (Piriyapongpipat *et al.*, 2024).

Evaluation frameworks designed specifically for learning ontologies include ODELO, an ontology-driven model for evaluating learning ontologies against pedagogical and technical criteria simultaneously (Kanellopoulos, 2009), and a family of ergonomic-metrics proposals that treat an educational ontology's usability for instructors and learners as a first-class quality dimension, both in its original formulation (Gavrilova *et al.*, 2012) and in later work on the cognitive ergonomics of educational ontologies (Popova, 2020). Complementary work has proposed dedicated quality frameworks for instructional-design ontologies, arguing that instructional-design artefacts require criteria beyond those used for descriptive domain ontologies (Li & Amorim, 2010), while other studies have examined how multilingualism interacts with ontology evaluation in educational contexts where course content is delivered in more than one language (Ivanova & Popov, 2020).

Several national- and institution-level efforts illustrate education-ontology evaluation at scale: the OntoSINAES methodology develops an ontology network for Brazil's national higher-education evaluation system (Sila *et al.*, 2018), and Eduonto proposes an ontology specifically for educational assessment processes (Marzano & Notti, 2015). Related work on ontologies for the evaluative process in higher education has been reported in the Spanish-language literature (Tabares García & Jiménez Builes, 2014), and a Portuguese-language literature review surveys ontologies applied to e-learning more broadly, corroborating the observation that evaluation remains the least standardized stage of the ontology life cycle in this domain (Nunes *et al.*, 2018). Taken together, this body of work establishes that education-domain ontologies must be assessed not only on the general structural, logical, and usability dimensions identified in Section 2, but also on domain-specific pedagogical-fitness criteria

(taxonomy alignment, objective traceability, and curriculum granularity) that have so far been addressed only piecemeal, motivating the unified formal methodology proposed in Section 4.

4. Proposed Formal Evaluation Methodology

This section presents the proposed methodology for evaluating education-domain ontologies and the knowledge graphs instantiated from them. It responds to the gap identified in Sections 2 and 3: existing frameworks supply general-purpose criteria and metrics, but none offers education-specific formulas, a reproducible composite score, or an auditable stakeholder-weighting protocol. The section is organized in four parts. Section 4.1 introduces the formal setting and the composite quality score. Section 4.2 gives an explicit, normalized definition for each of the five evaluation layers. Section 4.3 describes the protocol used to elicit the layer weights from stakeholders. Section 4.4 summarizes the steps for applying the methodology in practice.

4.1 Formal Setting and Composite Score

We model an education-domain ontology as a tuple $O = (C, R, A, I)$, where C is the set of classes (e.g., Syllabus, LearningObjective, Assessment, Tool, Learner), R is the set of object and data properties relating classes, A is the set of logical axioms (subsumption, disjointness, cardinality, and equivalence statements), and I is the set of individuals. A knowledge graph instantiated from O is a labelled graph $G = (V, E, L)$ where vertices V correspond to individuals in I, edges E correspond to asserted or inferred relations drawn from R, and L is a labelling function mapping vertices and edges back to classes and properties in O. Evaluation is defined over the pair (O, G): the ontology is assessed on its own terms, and the knowledge graph is assessed for how faithfully and completely it instantiates the ontology. Graph-theoretic descriptors of G are reported with every evaluation so that its size and connectivity are explicit: the number of vertices $|V|$ and edges $|E|$, the mean out-degree $\frac{|E|}{|V|}$, and the density $\frac{|E|}{|V|(|V|-1)}$. This follows the descriptive graph-modelling practice used in recent NIJOR studies (Babarinsa *et al.*, 2026; Ojo & Babarinsa, 2026).

The methodology decomposes ontology quality into five layers: a Structural layer $S(O)$; a Logical/Semantic layer $L(O)$; a Functional layer $F(O, G)$; a Pedagogical-Fitness layer $P(O, G)$, specific to the education domain; and a Usability/Social layer $U(O)$. Each layer is a score in the interval $[0,1]$, with higher values indicating better quality, and all sub-scores from which the layers are built are normalized to $[0,1]$ as well. Section 4.2 gives an explicit formula for every layer.

The overall ontology-quality score $\mathrm{OQ}(O, G)$ is computed as a weighted sum of the five layer scores using Eq. (1)

$$\mathrm{OQ}(O, G) = w_1 \cdot S(O) + w_2 \cdot L(O) + w_3 \cdot F(O, G) + w_4 \cdot P(O, G) + w_5 \cdot U(O) \qquad (1)$$

with $\sum_i w_i = 1$ and each $w_i \geq 0$, so that $\mathrm{OQ}(O, G)$ also lies in $[0,1]$. The weights are not fixed a priori; they are elicited from stakeholders by the protocol of Section 4.3. The layer scores are defined next.

4.2 Layer Definitions

Each layer score is a convex combination of normalized sub-scores $s \in [0,1]$: the coefficients are non-negative and sum to one within the layer and, unless the stakeholder panel decides otherwise, are equal.

4.2.1 Structural layer

$S(O)$ Captures class/relation richness, taxonomic depth, and structural redundancy or tangledness, consistent with the metric families used in OQuaRE-based evaluation (Duque-Ramos *et al.*, 2014; Duque-Ramos *et al.*, 2013; Lourdusamy & John, 2018). Let $H \subseteq A$ be the set of asserted direct

subsumption axioms, d_{avg} the mean depth of the leaf classes of the taxonomy, $[d_{\min}, d_{\max}]$ the depth band judged appropriate by the panel, $C_{\text{multi}} \subseteq C$ the classes with more than one asserted direct superclass, and $C_{\text{dup}} \subseteq C$ the classes that duplicate another class (same label or asserted equivalence). Then

$$s_{\text{rich}} = \frac{|R|}{|R| + |H|} \tag{2}$$

$$s_{\text{depth}} = \begin{cases} 1, & d_{\min} \le d_{\text{avg}} \le d_{\max} \\ \max\left\{0,\ 1 - \dfrac{\delta}{d_{\max}}\right\}, & \text{otherwise} \end{cases} \tag{3}$$

with $\delta = \min(|d_{\text{avg}} - d_{\min}|, |d_{\text{avg}} - d_{\max}|)$.

$$s_{\text{tang}} = 1 - \frac{|C_{\text{multi}}|}{|C|}, \qquad s_{\text{dup}} = 1 - \frac{|C_{\text{dup}}|}{|C|} \tag{4}$$

$$S(O) = \frac{s_{\text{rich}} + s_{\text{depth}} + s_{\text{tang}} + s_{\text{dup}}}{4} \tag{5}$$

4.2.2 Logical/semantic layer

$L(O)$ captures consistency, satisfiability, and axiom correctness as verified by a description-logic reasoner, in the spirit of modular evaluation systems such as Delta (Kondylakis *et al.*, 2021). Let $\chi(O) = 1$ if the reasoner reports O consistent and 0 otherwise, $C_{\perp} \subseteq C$ the unsatisfiable classes, and $A_{\text{bad}} \subseteq A$ the axioms flagged as redundant or erroneous by the reasoner or by manual review. Then

$$L(O) = \chi(O) \cdot \left[\frac{1}{2}\left(1 - \frac{|C_{\perp}|}{|C|}\right) + \frac{1}{2}\left(1 - \frac{|A_{\text{bad}}|}{|A|}\right)\right] \tag{6}$$

An inconsistent ontology therefore scores 0 on this layer, because none of its other logical properties can be relied upon.

4.2.3 Functional layer

$F(O, G)$ is the proportion of competency questions that the ontology and its knowledge graph answer correctly, following requirement-oriented practice (Grüninger & Fox, 1995; Yu *et al.*, 2009) and the goal–question–metric logic of FOCA (Bandeira *et al.*, 2016). Let Q be the set of competency questions (CQs), each formalized as a SPARQL 1.1 query (Harris & Seaborne, 2013) with a gold answer set $\text{Gold}(q)$ agreed by domain experts, and let $\text{Ans}_G(q)$ be the answers returned by evaluating the query over G. With precision $P_q = \frac{|\text{Ans}_G(q) \cap \text{Gold}(q)|}{|\text{Ans}_G(q)|}$, recall $R_q = \frac{|\text{Ans}_G(q) \cap \text{Gold}(q)|}{|\text{Gold}(q)|}$ (both set to 0 when $\text{Ans}_G(q)$ is empty), and $\text{F1}_q = \frac{2P_qR_q}{P_q + R_q}$, a question passes when $\text{F1}_q \ge \tau$; the default is $\tau = 1$, i.e., an exact match. Then

$$F(O, G) = \frac{1}{|Q|} \sum_{q \in Q} \mathbf{1}\left[\text{F1}_q \ge \tau\right] \tag{7}$$

The mean of P_q and R_q over Q is reported alongside $F(O, G)$ as a secondary diagnostic.

4.2.4 Pedagogical-fitness layer

$P(O, G)$, specific to the education domain, draws on ergonomic and instructional-design quality criteria (Gavrilova *et al.*, 2012; Popova, 2020; Li & Amorim, 2010) and on Bloom's taxonomy as revised by Anderson and Krathwohl (Bloom et al., 1956; Anderson & Krathwohl, 2001). Let Obj be the set of learning objectives in G, Asm the set of assessment items, and $B =$

{Remember, Understand, Apply, Analyze, Evaluate, Create} the six cognitive-process levels. For an objective o, let $\text{units}(o)$ be the number of separately assessable sub-outcomes linked to it; $[g_{\min}, g_{\max}]$ is the grain-size band the panel considers suitable for adaptive path generation. The three sub-scores are

$$s_{\text{trace}} = \frac{|\{o \in \text{Obj}:\ o \text{ carries exactly one level in } B\}|}{|\text{Obj}|} \tag{8}$$

$$s_{\text{align}} = \frac{|\{a \in \text{Asm}:\ a \text{ measures some } o \in \text{Obj},\ \text{link confirmed in expert review}\}|}{|\text{Asm}|} \tag{9}$$

$$s_{\text{gran}} = \frac{|\{o \in \text{Obj}:\ g_{\min} \le \text{units}(o) \le g_{\max}\}|}{|\text{Obj}|} \tag{10}$$

$$P(O,G) = \mu_1 s_{\text{trace}} + \mu_2 s_{\text{align}} + \mu_3 s_{\text{gran}}, \qquad \mu_1 + \mu_2 + \mu_3 = 1 \tag{11}$$

The default is $\mu_1 = \mu_2 = \mu_3 = 1/3$. Because s_{align} involves judgement, link confirmation is carried out by at least two reviewers and their agreement is reported with Cohen's kappa (Cohen, 1960).

4.2.5 Usability/social layer

$U(O)$ captures documentation quality, reuse potential, and community adoption, following the social-factors literature on ontology reuse (Talebpour *et al.*, 2020; Talebpour, 2020). Let $E = C \cup R$, $E_{\text{doc}} \subseteq E$ the entities that carry both an rdfs:label and an rdfs:comment annotation, $E_{\text{ext}} \subseteq E$ the entities reused from published external vocabularies, n_{use} the number of documented downstream uses (imports, citations, or deployments), and n^* the adoption level that the panel regards as full adoption. Then

$$s_{\text{doc}} = \frac{|E_{\text{doc}}|}{|E|}, \qquad s_{\text{reuse}} = \frac{|E_{\text{ext}}|}{|E|}, \qquad s_{\text{adopt}} = \min\left\{1, \frac{n_{\text{use}}}{n^*}\right\} \tag{12}$$

$$U(O) = \nu_1 s_{\text{doc}} + \nu_2 s_{\text{reuse}} + \nu_3 s_{\text{adopt}}, \qquad \nu_1 + \nu_2 + \nu_3 = 1 \tag{13}$$

The default is $\nu_1 = \nu_2 = \nu_3 = 1/3$. Taken together, Eqs. (2)–(13) make every layer a function of counts that can be obtained from the ontology, the reasoner, the query log, and the review record, which is what allows a third party to repeat the evaluation.

4.3 Weight Elicitation Protocol

Rather than fixing the weights a priori, the methodology elicits them from a panel of K stakeholders (instructors, instructional designers, and system engineers; we recommend $K \ge 5$ with every group represented) through an iterative procedure modelled on ROMEO (Yu *et al.*, 2009) and on OntoElect's saturation- and vote-based fitness quantification (Tatarintseva *et al.*, 2013b; Ermolayev, 2018; Vasileyko & Ermolayev, 2018). The mechanics are as follows.

(a) *Voting.* In round $r = 1,2,\ldots$, after reviewing the layer definitions and a sample of competency questions, each stakeholder k allocates one unit of weight across the five layers, giving a vote vector $v_k(r) = (v_{k1}, \ldots, v_{k5})$ with $v_{ki} \ge 0$ and $\sum_i v_{ki} = 1$. From round 2 onward, stakeholders see the previous round's aggregate before voting again.

(b) *Aggregation.* The panel weight vector of round r is the arithmetic mean of the votes,

$$w_i(r) = \frac{1}{K}\sum_{k=1}^{K} v_{ki}(r), \qquad i = 1,\ldots,5 \tag{14}$$

so that $\sum_i w_i\,(r) = 1$ holds automatically. Where stakeholders prefer pairwise judgements to point allocation, the alternative is the analytic hierarchy process: Saaty-scale comparison matrices are aggregated across experts by the geometric mean, weights are taken from the principal eigenvector, and the judgements are accepted only if the consistency ratio is below 0.10, as in the multi-criteria study (Barma *et al.*, 2026).

(c) *Saturation (stopping rule).* Define the round-to-round change $\Delta(r) = \max_i |w_i(r) - w_i(r-1)|$ and let W(r) be Kendall's coefficient of concordance of the stakeholders' rankings of the five layers (Kendall & Babington Smith, 1939). Elicitation stops when $\Delta(r) \le \varepsilon$ in two consecutive rounds and $W(r) \ge 0.70$. The recommended defaults are $\varepsilon = 0.02$ and a maximum of $R_{\max} = 5$ rounds; if $R_{\max}$ is reached without saturation, the weights of round $R_{\max}$ are used and the absence of saturation is reported.

(d) *Final weights and robustness.* The final weights are $w = w(R)$, where R is the stopping round. To show that conclusions do not hinge on the exact weights, each evaluation also reports OQ under equal weights and under perturbations of ± 0.05 of each weight (renormalized so that the weights sum to one). Every elicitation is documented in a record table (rounds, K, aggregated weights, Δ, and W; see Table 4) so that the weights can be audited.

4.4 Application Steps

The methodology is applied in five steps: (1) requirements elicitation, in which competency questions and stakeholder priorities are gathered; (2) layer-wise metric selection, in which the sub-scores of Eqs. (2)–(13) are instantiated and their parameters (the depth band $[d_{\min}, d_{\max}]$, the granularity band $[g_{\min}, g_{\max}]$, and the adoption level n^*) are fixed, drawing on the criteria-selection literature (Hooi *et al.*, 2015; Kim & Oh, 2019); (3) weight elicitation, using the stakeholder-voting protocol of Section 4.3; (4) scoring, combining automated metric computation with expert-in-the-loop review for the pedagogical-fitness layer, consistent with expert-informed evaluation practice (Rodriguez Apodaca *et al.*, 2019); and (5) iterative refinement, in which low-scoring layers are traced back to specific ontology-design decisions and revised. Table 1 summarizes the five layers, representative criteria, and the metrics defined in Section 4.2.

Table 1: Layers, criteria, and metrics of the proposed formal evaluation methodology. All sub-scores are normalized to the interval [0,1]

Layer	Representative Criteria	Representative Metrics	Key Source(s)
Structural, $S(O)$ — Eq. (5)	Class/relation richness; taxonomic depth & breadth; redundancy	Relationship richness; depth-band fit; 1 − tangledness; 1 − duplication	(Duque-Ramos *et al.*, 2014; Duque-Ramos *et al.*, 2013; Lourdusamy & John, 2018)
Logical/Semantic, $L(O)$ — Eq. (6)	Consistency; satisfiability; axiom correctness	Consistency indicator × (1 − unsatisfiable-class share; 1 − faulty-axiom share)	(Kondylakis *et al.*, 2021)
Functional, F(O,G) — Eq. (7)	Competency-question coverage; task fitness	CQ pass rate over SPARQL-formalized CQs; precision/recall on G	(Bandeira *et al.*, 2016; Grüninger & Fox, 1995; Yu *et al.*, 2009)
Pedagogical Fitness, P(O,G) — Eq. (11)	Taxonomy-level traceability; objective–assessment alignment; curriculum granularity	Taxonomy traceability; objective–assessment alignment; granularity-band share	(Gavrilova *et al.*, 2012; Popova, 2020; Li & Amorim, 2010; Anderson & Krathwohl, 2001)
Usability/Social, $U(O)$ — Eq. (13)	Documentation quality; reuse potential; community adoption	Annotation completeness; external-term reuse ratio; adoption level	(Talebpour *et al.*, 2020; Talebpour, 2020)

Note. All sub-scores and layer scores are normalized to $[0,1]$; higher values indicate better quality. Abbreviations: CQ = competency question; Bloom/Anderson = Bloom's taxonomy of educational objectives (Bloom et al., 1956) as revised by Anderson and Krathwohl (2001); OQuaRE = Ontology Quality Requirements and Evaluation; ROMEO = Requirements-

Oriented Methodology for Evaluating Ontologies; GQM = Goal–Question–Metric; SPARQL = SPARQL Protocol and RDF Query Language; O = ontology; G = knowledge graph.

5. Applying the Methodology to an Education-Domain Knowledge Graph

To illustrate the methodology, we apply it to a representative education-domain knowledge graph modelling a course syllabus, its learning objectives, its supporting tools and assignments, and the assessment scores recorded against each objective, the same class of ontology used to drive knowledge-graph-based learning-analytics systems built on taxonomy-aligned assignment data. The ontology defines classes for Syllabus, Subject, Objective, Support Tool, Learning Activity Tool, Learning Object, and Score, related by subsumption and object-property axioms (e.g., Subject hasObjective Objective; Objective isSupportedBy SupportTool), broadly consistent with the class structures reported for syllabus- and curriculum-oriented education ontologies (Ilkou *et al.*, 2021; Kanellopoulos, 2009). The graph G was built in RDF/Turtle using Protégé 5.6 and Apache Jena and has the dimensions reported in Table 2; all counts in Table 2 are the inputs to Eqs. (2)–(13).

Table 2: Dimensions of the evaluated ontology O and knowledge graph G

Quantity	Symbol	Value
Classes	$\|C\|$	15
Properties (object and datatype)	$\|R\|$	11
Logical axioms	$\|A\|$	50
of which direct subsumption axioms	$\|H\|$	14
Individuals = graph vertices	$\|I\| = \|V\|$	320
Graph edges (asserted + inferred)	$\|E\|$	1,184
Mean out-degree	$\frac{\|E\|}{\|V\|}$	3.70
Density	$\frac{\|E\|}{\|V\|(\|V\|-1)}$	0.0116
Learning objectives; assessment items	$\|\text{Obj}\|$; $\|\text{Asm}\|$	48; 64
Competency questions	$\|Q\|$	25
Reasoner and version	—	HermiT 1.4.3.456

Applying Step 2 of the methodology, we instantiated the sub-scores of Eqs. (2)–(13) and computed the layer scores summarized in Table 3. The Structural layer scored highly because the class hierarchy is shallow and free of cycles, a property favoured by OQuaRE-style structural metrics (Duque-Ramos *et al.*, 2014). The Logical layer also scored highly, since reasoner checking reported no unsatisfiable classes. The Functional layer scored moderately: several competency questions concerning cross-subject prerequisite relationships could not be answered directly from the graph, indicating a missing object property. Curriculum ontologies in recent work encode topic and course continuation explicitly through a dedicated object property (Piriyapongpipat *et al.*, 2024). The Pedagogical-Fitness layer, the layer specific to this methodology, scored lowest: only a subset of learning objectives carried an explicit Bloom/Anderson taxonomy-level annotation, and a smaller subset of assessment items were linked back to the specific objective they were intended to measure, echoing the traceability gaps noted in instructional-design-oriented ontology-quality work (Li & Amorim, 2010). The Usability layer scored moderately, reflecting adequate but incomplete class and property documentation.

Functional layer in detail. The competency-question set Q contained 25 questions, elicited from instructors' syllabus-planning requirements and the data needs of the adaptive-path generator. Each

question was formalized as a SPARQL 1.1 query and evaluated over G against a gold answer set agreed by three domain experts (two course instructors and one instructional designer) (Eq. (7) with $\tau = 1$). 17 of the 25 questions passed, so $F(O,G) = 17/25 = 0.68$. The failing questions were those concerning cross-subject prerequisites (CQ3 in Appendix A), which return an empty answer because the ontology defines no prerequisite property. Three representative queries are listed in Appendix A.

Table 3: Illustrative layer-wise evaluation scores for the case-study syllabus/learning-objective knowledge graph. All scores are normalized to $[0,1]$

Layer	Normalized score $[0,1]$	Interpretation / Recommended Action
Structural, $S(O)$	0.86	Hierarchy is shallow and acyclic; no action required.
Logical/Semantic, $L(O)$	0.93	No unsatisfiable classes; minor redundant axioms flagged for cleanup.
Functional, F(O,G)	0.68	Add a hasPrerequisite object property to answer cross-subject CQs.
Pedagogical Fitness, P(O,G)	0.52	Complete Bloom/Anderson-level annotation for all objectives; link remaining assessment items to their objectives.
Usability/Social, $U(O)$	0.71	Extend rdfs:comment annotations on Support Tool and Learning Object classes.

Note. Scores are normalized to $[0,1]$; higher is better. S, L, F, P, and U are defined in Eqs. (2)–(13). CQ = competency question; Bloom/Anderson = Bloom's taxonomy as revised by Anderson and Krathwohl; rdfs:comment = RDF Schema comment annotation.

With equal weights ($w_i = 0.2$ for all i), Eq. (1) gives $\mathrm{OQ}(O,G) = 0.2 \times (0.86 + 0.93 + 0.68 + 0.52 + 0.71) = 0.74$. Under a deployment-specific weighting elicited from the stakeholder panel by the protocol of Section 4.3, which assigned a higher weight to pedagogical fitness ($w_4 = 0.35$) because the target system is an adaptive-path generator, the composite score falls to $\mathrm{OQ}(O,G) = 0.69$, correctly surfacing the pedagogical-traceability gap as the priority for the next revision cycle rather than averaging it away. Table 4 documents the elicitation: the panel comprised 7 stakeholders (three instructors, two instructional designers, and two system engineers); saturation ($\Delta(r) \leq 0.02$ in two consecutive rounds and $W(r) \geq 0.70$) was reached after 3 rounds; and the final weights are $w = (0.15, 0.15, 0.20, 0.35, 0.15)$, which sum to one. Perturbing each weight by ± 0.05 and renormalizing changed $\mathrm{OQ}(O,G)$ by at most 0.013, so the ordering of the layers as improvement priorities is unchanged. This behaviour is the intended effect of stakeholder-elicited weighting: two ontologies with the same average quality can be ranked differently once weights reflect what actually matters for the deployment context, consistent with the requirements-oriented evaluation philosophy of ROMEO (Yu *et al.*, 2009) and OntoElect (Ermolayev, 2018).

Table 4: Weight-elicitation record for the case study (aggregated panel weights per round, Eq. (14)

Round r	K	w_1 (S)	w_2 (L)	w_3 (F)	w_4 (P)	w_5 (U)	Δ(r)	W(r)
1	7	0.16	0.16	0.20	0.33	0.15	—	0.61
2	7	0.15	0.16	0.20	0.34	0.15	0.01	0.69
Final $(R = 3)$	7	0.15	0.15	0.20	0.35	0.15	0.01	0.74

6. Conclusion and Future Research

This paper proposed a formal, multi-layer methodology for evaluating education-domain ontologies and the knowledge graphs derived from them. By combining structural, logical, functional, and usability layers already established in the general ontology-evaluation literature with a new

pedagogical-fitness layer specific to instructional content (taxonomy-level traceability, objective–assessment alignment, and curriculum granularity), the methodology closes a gap left by general-purpose frameworks such as OQuaRE, FOCA, and Delta, which do not explicitly assess these pedagogical properties. The weighted, stakeholder-elicited scoring model (Eqs. (1)–(14) allows the same evaluation instrument to be reused across deployment contexts (course catalogues, recommendation systems, adaptive-path generators) while still producing a reproducible, criterion-based score, and the case study demonstrates that the pedagogical-fitness layer surfaces gaps that structural and logical metrics alone would miss.

Limitations. The case study applies the methodology to a single knowledge graph and a single stakeholder panel, so the elicited weights illustrate the protocol rather than establish universal values. The alignment sub-score s_{align} depends on expert judgement, and the depth band, granularity band, and adoption level are panel-defined parameters that must be reported with every evaluation. The aggregation is a linear weighted sum, which allows a high score on one layer to compensate for a low score on another; the per-layer scores in Table 3 are therefore always reported alongside the composite.

Future work includes automating the computation of the structural and logical layers through a reasoner-integrated pipeline, extending the functional-layer evaluation with a standard competency-question benchmark for the education domain (cf. Lippolis *et al.*, 2025a) and with LLM-assisted competency-question verification (Lippolis *et al.*, 2025b), and validating the methodology empirically across multiple published education ontologies, including Bowlogna, AIISO, and EduCOR-style ontologies. We also plan to study how the multilingualism issues identified in prior work (Ivanova & Popov, 2020) interact with the pedagogical-fitness layer when course content spans more than one language, and to run the stakeholder weight-elicitation process with a larger and more diverse panel of instructors and instructional designers to assess how stable the resulting weights are across institutions. Finally, we plan to compare the arithmetic-mean aggregation of Section 4.3 with AHP-based and fuzzy aggregation schemes of the kind used in decision-analytics studies (Barma *et al.*, 2026).

Author Contributions

Conceptualization, Manunyi Godfrey and Kayode Oladipupo Oluborode; methodology, Manunyi Godfrey and Shaphat Safeton; software, Shaphat Safeton; validation, Manunyi Godfrey, Shaphat Safeton and Kayode Oladipupo Oluborode; formal analysis, Manunyi Godfrey and Shaphat Safeton; investigation, Manunyi Godfrey and Shaphat Safeton; resources, Shaphat Safeton; data curation, Shaphat Safeton; writing (original draft preparation), Manunyi Godfrey; writing (review and editing), Manunyi Godfrey, Shaphat Safeton and Kayode Oladipupo Oluborode; visualization, Manunyi Godfrey and Shaphat Safeton; supervision, Kayode Oladipupo Oluborode; project administration, Manunyi Godfrey. All authors have read and agreed to the published version of the manuscript.

Funding

This research received no external funding.

Data Availability Statement

The data supporting the reported results are the case-study ontology, the knowledge graph instantiated from it (RDF/Turtle serialization), the 25 competency questions with their SPARQL 1.1 formalizations and expert-agreed gold answer sets, the reasoner output, and the weight-elicitation record. Representative competency questions and queries are given in Appendix A, and the graph dimensions are reported in Table 2. The complete files are available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgement

The authors sincerely thank Dr. Oloborade Kayode of the Department of Computer Science for his encouragement and support, which were instrumental in bringing this research to fruition. This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

Amirhosseini, M., & Salim, J. (2019). A synthesis survey of ontology evaluation tools, applications and methods to propose a novel branch in evaluating the structure of ontologies: Graph-independent approach. *International Journal of Computer, 33*(1), 46–68.

Anderson, L. W., & Krathwohl, D. R. (Eds.). (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives* (Complete ed.). Addison Wesley Longman.

Babarinsa, O., Edime, G., & Dare, E. (2026). A graph analysis of topological structure in the Bus Rapid Transit transport network of Lagos State. *Nigerian Journal of Operations Research, 3*(2), 73–81. https://doi.org/10.67868/nijor/0302.007

Bandeira, J., Bittencourt, I. I., Espinheira, P., & Isotani, S. (2016). *FOCA: A methodology for ontology evaluation* (arXiv:1612.03353). arXiv. https://doi.org/10.48550/arXiv.1612.03353

Barma, M., Babando, A. H., & Emmanuel, J. J. (2026). Multi-criterion decision-making approach to prioritize strategies for combating smuggling and border control along the North-east borders of Nigeria. *Nigerian Journal of Operations Research, 3*(2), 56–72. https://doi.org/10.67868/nijor/0302.006

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain* (207 pp.). New York, NY: David McKay Company.

Bouiadjra, A. B., & Benslimane, S. M. (2011a). FOEval: Full ontology evaluation. In *Proceedings of the IEEE International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE)*. IEEE. https://doi.org/10.1109/NLPKE.2011.6138244

Bouiadjra, A. B., & Benslimane, S. M. (2011b). Ontology evaluation: State of the art, new approach and perspectives. In *Proceedings of the International Conference on Knowledge Engineering and Ontology Development (KEOD 2011)* (pp. 365–368). SciTePress. https://doi.org/10.5220/0003635003650368

Cardinale, Y., Cornejo-Lupa, M. A., Ticona-Herrera, R., et al. (2020). A methodological approach to compare ontologies: Proposal and application for SLAM ontologies. In *Proceedings of the International Conference on Information Integration and Web-based Applications & Services (iiWAS 2020)*. ACM. https://doi.org/10.1145/3428757.3429091

Cardinale, Y., Cornejo-Lupa, M. A., Pinto-De la Gala, A., & Ticona-Herrera, R. (2021). Application of a methodological approach to compare ontologies. *International Journal of Web Information Systems, 17*(4), 333–376. https://doi.org/10.1108/IJWIS-03-2021-0036

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46.

Debnath, N. C., & Patel, A. (2023). Ontology evaluation tools: Current and future research. *Recent Advances in Computer Science and Communications, 16*(6). https://doi.org/10.2174/2666255815666220411081837

Degbelo, A. (2017). A snapshot of ontology evaluation criteria and strategies. In *Proceedings of SEMANTiCS 2017*. ACM. https://doi.org/10.1145/3132218.3132219

Duque-Ramos, A., Fernández-Breis, J. T., Iniesta, M., et al. (2013). Evaluation of the OQuaRE framework for ontology quality. *Expert Systems with Applications, 40*(7), 2459–2468. https://doi.org/10.1016/J.ESWA.2012.11.004

Duque-Ramos, A., Boeker, M., Jansen, L., et al. (2014). Evaluating the Good Ontology Design guideline (GoodOD) with the ontology quality requirements and evaluation method and metrics (OQuaRE). *PLOS ONE, 9*(8), e104463. https://doi.org/10.1371/JOURNAL.PONE.0104463

Ermolayev, V. (2018). OntoElecting requirements for domain ontologies: The case of time domain. *Enterprise Modelling and Information Systems Architectures (EMISAJ), 13*, 86–109. https://doi.org/10.18417/EMISA.SI.HCM.9

Fathian Dastgerdi, A. (2012). Ontology evaluation: Consideration of criteria, approaches and layers. *Iranian Journal of Information Processing and Management, 27*(2), 533–559. https://jipm.irandoc.ac.ir/article_699141.html

García, J. L., García-Peñalvo, F. J., & Therón, R. (2010). A survey on ontology metrics. In *Knowledge Management, Information Systems, E-Learning, and Sustainability Research* (Communications in Computer and Information Science, Vol. 111). Springer. https://doi.org/10.1007/978-3-642-16318-0_4

Gavrilova, T., Gorovoy, V., & Bolotnikov, E. (2012). New ergonomic metrics for educational ontology design and evaluation. In *Frontiers in Artificial Intelligence and Applications*. IOS Press. https://doi.org/10.3233/978-1-61499-125-0-361

Gelernter, J., & Jha, J. (2016). Challenges in ontology evaluation. *Journal of Data and Information Quality, 7*(3), 1–4. https://doi.org/10.1145/2935751

Gómez-Pérez, A. (1996). Towards a framework to verify knowledge sharing technology. *Expert Systems with Applications, 11*(4), 519–529. https://doi.org/10.1016/S0957-4174(96)00067-X

Grüninger, M., & Fox, M. S. (1995). Methodology for the design and evaluation of ontologies. In *Proceedings of the Workshop on Basic Ontological Issues in Knowledge Sharing, held in conjunction with IJCAI-95* (pp. 1–10). Montréal, Canada.

Harris, S., & Seaborne, A. (Eds.). (2013). *SPARQL 1.1 query language* (W3C Recommendation). World Wide Web Consortium. https://www.w3.org/TR/sparql11-query/

Heiyanthuduwage, S. R. (2022). A review: Status quo and current trends in e-learning ontologies. Springer. https://doi.org/10.1007/978-3-030-93904-5_12

Hicks, A. (2017). Metrics and methods for comparative ontology evaluation. *Ciência da Informação, 46*(1). https://doi.org/10.18225/ci.inf.v46i1.4012

Hooi, Y. K., Hassan, M. F., & Shariff, A. M. (2015). Ontology evaluation: A criteria selection framework. In *Proceedings of the International Symposium on Mathematical Sciences and Computing Research (iSMSC)*. IEEE. https://doi.org/10.1109/ISMSC.2015.7594069

Ilkou, E., Abu-Rasheed, H., Tavakoli, M., Hakimov, S., Kismihók, G., Auer, S., & Nejdl, W. (2021). EduCOR: An educational and career-oriented recommendation ontology. In *The Semantic Web – ISWC 2021* (Lecture Notes in Computer Science, Vol. 12922). Springer. https://doi.org/10.1007/978-3-030-88361-4_32

Ivanova, T., & Popov, M. (2020). Ontology evaluation and multilingualism. In *Proceedings of CompSysTech '20*. ACM. https://doi.org/10.1145/3407982.3407989

Kanellopoulos, D. N. (2009). ODELO: An ontology-driven model for the evaluation of learning ontologies. *International Journal of Learning Technology, 4*(1/2). https://doi.org/10.1504/IJLT.2009.024717

Kendall, M. G., & Babington Smith, B. (1939). The problem of m rankings. *The Annals of Mathematical Statistics, 10*(3), 275–287.

Khan, Z. C., & Keet, C. M. (2016). Dependencies between modularity metrics towards improved modules. In *Knowledge Engineering and Knowledge Management (EKAW 2016)* (Lecture Notes in Computer Science, Vol. 10024, pp. 400–415). Springer.

Khoiruddin, M., Kusumawardani, S. S., Hidayah, I., et al. (2023). A review of ontology development in the e-learning domain: Methods, roles, evaluation. In *Proceedings of IC3INA 2023*. IEEE. https://doi.org/10.1109/ic3ina60834.2023.10285789

Kim, S., & Oh, S. G. (2019). Extracting and applying evaluation criteria for ontology quality assessment. *Library Hi Tech*. https://doi.org/10.1108/LHT-01-2019-0012

Kondylakis, H., Nikolaos, A., Dimitra, P., et al. (2021). Delta: A modular ontology evaluation system. *Information, 12*(8), 301. https://doi.org/10.3390/INFO12080301

Lei, X. (2016). Research advances in ontology evaluation. *Journal of the China Society for Scientific and Technical Information*. https://doi.org/10.3772/j.issn.1000-0135.2016.007.011

Li, X., & Amorim, R. (2010). Development of a framework for quality instructional design ontologies. In *Anais do Simpósio Brasileiro de Informática na Educação (SBIE 2010)*.

Lippolis, A. S., Ragagni, M. D., Ciancarini, P., Nuzzolese, A. G., & Presutti, V. (2025a). *Bench4KE: Benchmarking automated competency question generation* (arXiv:2505.24554v3). arXiv. https://doi.org/10.48550/arXiv.2505.24554

Lippolis, A. S., Saeedizade, M. J., Keskisärkkä, R., Gangemi, A., Blomqvist, E., & Nuzzolese, A. G. (2025b). Large language models assisting ontology evaluation. In D. Garijo et al. (Eds.), *The Semantic Web – ISWC 2025* (Lecture Notes in Computer Science, Vol. 16140, pp. 502–520). Springer. https://doi.org/10.1007/978-3-032-09527-5_27

López Rodríguez, Y. A., Hidalgo-Delgado, Y., & Silega Martínez, N. (2018). Un método práctico para la evaluación de ontologías. In *Proceedings of the International Workshop on Semantic Web (IWSW 2018)* (CEUR Workshop Proceedings, Vol. 2096, pp. 76–85). CEUR-WS.org. https://ceur-ws.org/Vol-2096/

Lourdusamy, R., & John, A. (2017). A survey on methods and tools for ontology evaluation. https://doi.org/10.5958/0975-8089.2017.00015.X

Lourdusamy, R., & John, A. (2018). A review on metrics for ontology evaluation. In *2018 2nd International Conference on Inventive Systems and Control (ICISC)*. IEEE. https://doi.org/10.1109/ICISC.2018.8399041

Marzano, A., & Notti, A. M. (2015). Eduonto: An ontology for educational assessment. https://doi.org/10.20368/1971-8829/978

McDaniel, M., & Storey, V. C. (2019). Evaluating domain ontologies: Clarification, classification, and challenges. *ACM Computing Surveys, 52*(4). https://doi.org/10.1145/3329124

Mohammed, S., Ehrlinger, L., Harmouch, H., Naumann, F., & Srivastava, D. (2025). The five facets of data quality assessment. *SIGMOD Record, 54*, 18–27. https://doi.org/10.1145/3749116.3749120

Nefzi, H., Farah, M., Farah, I. R., et al. (2014). A critical analysis of lifecycles and methods for ontology construction and evaluation. In *Proceedings of ATSIP 2014*. IEEE. https://doi.org/10.1109/ATSIP.2014.6834646

Norouzi, E., Waitelonis, J., & Sack, H. (2024). The landscape of ontologies in materials science and engineering: A survey and evaluation. In *Proceedings of the First International Workshop on Semantic Materials Science (SeMatS 2024)* (CEUR Workshop Proceedings, Vol. 3760, pp. 78–100). CEUR-WS.org. https://ceur-ws.org/Vol-3760/paper3.pdf

Novogrudska, R., & Popova, M. (2021). A comprehensive review of ontology-based information systems for educational process support. In *Proceedings of UkrMiCo 2021*. IEEE. https://doi.org/10.1109/ukrmico52950.2021.9716675

Nunes, R. C. T., Gonçalves, A. L., de Souza, J. A., et al. (2018). Uma revisão da literatura sobre ontologias aplicadas ao e-learning.

Ojo, O., & Babarinsa, O. (2026). Graph modelling of transaction networks for fraud detection and systemic risk assessment in Nigerian banks. *Nigerian Journal of Operations Research, 3*(2), 1–55. https://doi.org/10.67868/nijor/0302.005

Pellegrino, M. A., Rula, A., Ehrlinger, L., Micsik, A., Spahiu, B., & Etcheverry, L. (2026). Measuring what matters: User perceptions of knowledge graph quality dimensions in cultural heritage. In *Proceedings of the Workshop on Quality of Knowledge Graphs (QKG@ESWC 2026)* (CEUR Workshop Proceedings, Vol. 4205). CEUR-WS.org. https://ceur-ws.org/Vol-4205/paper14.pdf

Piriyapongpipat, P., Goldin, S., & Ditcharoen, N. (2024). An alternative approach to ontology-based curriculum development in higher education. *Smart Learning Environments, 11*(1), Article 20. https://doi.org/10.1186/s40561-024-00307-8

Popova, M. A. (2020). Cognitive ergonomics of educational ontology. https://doi.org/10.51707/2618-0529-2020-18-05

Rodriguez Apodaca, M. del C., Ramírez-Noriega, A., Martínez-Ramírez, Y., et al. (2019). Evaluation of ontologies based on experts knowledge. In *Proceedings of CONISOFT 2019*. IEEE. https://doi.org/10.1109/CONISOFT.2019.00028

Sanagavarapu, L. M., Gollapudi, S., Chimalakonda, S., et al. (2017). A lightweight approach for evaluating sufficiency of ontologies. In *Proceedings of SEKE 2017*. https://doi.org/10.18293/SEKE2017-185

Sila, C., Belo, O., & Barros, V. (2018). Methodology for the development of an ontology network on the Brazilian national system for the evaluation of higher education (OntoSINAES). https://doi.org/10.6084/m9.figshare.6388055

Silva-López, R. B., Méndez-Gurrola, I. I., & Pablo-Leyva, H. (2020). Comparative methodologies for evaluation of ontology design. Springer. https://doi.org/10.1007/978-3-030-60887-3_9

Tabares García, J. J., & Jiménez Builes, J. A. (2014). Ontología para el proceso evaluativo en la educación superior. *Revista Virtual Universidad Católica del Norte, 42*, 68–79.

Talebpour, M. (2020). *The evaluation of ontologies: Quality, reuse and social factors* [Doctoral thesis, Loughborough University]. https://doi.org/10.26174/thesis.lboro.12287273

Talebpour, M., Sykora, M. D., & Jackson, T. (2017). The role of community and social metrics in ontology evaluation: An interview study of ontology reuse. In *Proceedings of KEOD 2017* (pp. 119–127). SciTePress. https://doi.org/10.5220/0006589201190127

Talebpour, M., Sykora, M. D., & Jackson, T. (2020). The evaluation of ontologies for quality, suitability for reuse, and the significant role of social factors. Springer. https://doi.org/10.1007/978-3-030-49559-6_8

Tankeleviciene, L., & Damasevicius, R. (2009). Characteristics of domain ontologies for web based learning and their application for quality evaluation. *Informatics in Education, 8*. https://doi.org/10.15388/INFEDU.2009.09

Tartir, S., Arpinar, I. B., & Sheth, A. P. (2010). Ontological evaluation and validation. In *Theory and Applications of Ontology: Computer Applications* (pp. 115–130). Springer. https://doi.org/10.1007/978-90-481-8847-5_5

Tatarintseva, O., Borue, Y., & Ermolayev, V. (2013a). Validating OntoElect methodology in refining ICTERI scope ontology. In *ICT in Education, Research, and Industrial Applications* (Communications in Computer and Information Science). Springer. https://doi.org/10.1007/978-3-642-38370-0_12

Tatarintseva, O., Ermolayev, V., Keller, B., et al. (2013b). Quantifying ontology fitness in OntoElect using saturation- and vote-based metrics. In *ICT in Education, Research, and Industrial Applications* (Communications in Computer and Information Science). Springer. https://doi.org/10.1007/978-3-319-03998-5_8

Vasileyko, A., & Ermolayev, V. (2018). Evaluating the fitness of a domain ontology to formalized stakeholder requirements.

Wang, Y., & Wang, Y. (2021). A survey of ontologies and their applications in e-learning environments. *Journal of Web Engineering, 20*(6). https://doi.org/10.13052/JWE1540-9589.2061

Watsilla, H. (2026). Quality assurance as individual responsibility: The three-tier reality of research integrity in Nigerian public universities. *Nigerian Journal of Operations Research, 3*(2), 131–144. https://doi.org/10.67868/nijor/0302.011

Wilson, R. S. I., Goonetillake, J. S., et al. (2021). Analysis of ontology quality dimensions, criteria and metrics. In *Computational Science and Its Applications – ICCSA 2021*. Springer. https://doi.org/10.1007/978-3-030-86970-0_23

Wilson, R. S. I., Goonetillake, J. S., Indika, W. A., et al. (2023). A conceptual model for ontology quality assessment. *Semantic Web*. https://doi.org/10.3233/sw-233393

Xue, B., & Zou, L. (2023). Knowledge graph quality management: A comprehensive survey. *IEEE Transactions on Knowledge and Data Engineering, 35*(5). https://doi.org/10.1109/TKDE.2022.3150080

Yu, J., Thom, J. A., & Tam, A. (2009). Requirements-oriented methodology for evaluating ontologies. *Information Systems, 34*(8), 766–791.

Zouaq, A., & Nkambou, R. (2009). Evaluating the generation of domain ontologies in the Knowledge Puzzle project. *IEEE Transactions on Knowledge and Data Engineering, 21*(11), 1559–1572. https://doi.org/10.1109/TKDE.2009.25

Appendix A. Sample Competency Questions and SPARQL Queries

This appendix lists three of the 25 competency questions used for the Functional layer, with their SPARQL 1.1 formalizations (Harris & Seaborne, 2013), and one diagnostic query used for the Pedagogical-Fitness layer. The prefix edu: stands for the namespace of the evaluated ontology; the property names (hasObjective, isSupportedBy, recordedFor, hasPrerequisite, hasCognitiveLevel) are written as in the case-study ontology description and should be aligned with the actual IRIs of the evaluated ontology when the queries are re-run. A question passes when its result set matches the expert-agreed gold answer set ($\tau = 1$ in Eq. (7).

CQ1. Which learning objectives does each subject have, and which support tool supports each objective? *(Outcome: pass.)*

```
PREFIX edu: <https://example.org/edu#>
SELECT ?subject ?objective ?tool
WHERE {
  ?subject   a                 edu:Subject ;
             edu:hasObjective  ?objective .
  ?objective edu:isSupportedBy ?tool .
}
```

CQ2. Which scores have been recorded against each learning objective? *(Outcome: pass.)*

```
PREFIX edu: <https://example.org/edu#>
SELECT ?objective ?score
WHERE {
  ?score a edu:Score ;
         edu:recordedFor ?objective .
}
```

CQ3. Which subjects are prerequisites of which other subjects? *(Outcome: fail — the query returns an empty result because the ontology has no prerequisite property.)*

```
PREFIX edu: <https://example.org/edu#>
SELECT ?subject ?prerequisite
WHERE {
  ?subject a edu:Subject ;
           edu:hasPrerequisite ?prerequisite .
}
```

D1 (diagnostic for s_{trace}, Eq. (8). Which learning objectives carry no Bloom/Anderson level?

```
PREFIX edu: <https://example.org/edu#>
SELECT ?objective
WHERE {
  ?objective a edu:Objective .
  FILTER NOT EXISTS { ?objective edu:hasCognitiveLevel ?level . }
}
```