Toward a Comparison Score for the Open Research Knowledge Graph: Modeling Guideline and Metrics

Lena John[1], Sören Auer[1] and Oliver Karras[1]

[1] TIB - Leibniz Information Centre for Science and Technology, Hannover, Germany

Abstract

Research Knowledge Graphs (RKGs) have become a foundational technology for representing machine-actionable knowledge, enabling semantic search, automated reasoning, and knowledge discovery. However, the utility of a RKG is inherently bound to its quality. One established RKG is the Open Research Knowledge Graph (ORKG). It allows researchers to collaboratively create structured representations of scientific knowledge for so-called *Comparisons*, i.e., tabular state-of-the-art overviews that present structured knowledge across publications. As the volume of these community-generated Comparisons grows, ensuring their quality presents significant data management and curation challenges. To support consistent scientific knowledge representation, the ORKG Curation & Community Building (C&CB) team has established a modeling guideline that consolidates existing best practices and practical curation experience into recommendations for creating high-quality Comparisons. In this paper, we present a quality framework that translates these recommendations into quality metrics for assessing the extent to which ORKG Comparisons conform to the proposed modeling practices. The framework serves two complementary purposes. First, it enables continuous quality assessment through the existing curation dashboard *SciKGDash*, providing curators with insights into recurring modeling issues and the current state of ORKG Comparison quality. Second, it establishes the foundation for a future Comparison score that provides live feedback and actionable recommendations for contributors during the modeling process. As a first step toward this vision, we implement a preliminary subset of the proposed metrics in *SciKGDash* and apply them to ORKG Comparisons to establish a baseline assessment of current modeling practices. Our analysis reveals recurring quality issues, including empty Comparison cells, incorrectly assigned data types, and inconsistent predicate label formatting. These findings inform the development of a future quality model comprising quality (sub-)characteristics and measurement procedures that, together with the proposed recommendations and metrics, will establish the envisioned Comparison score.

Keywords

Research Knowledge Graph, Quality Management, Human-AI Collaboration, Crowdsourcing

CEUR-WS.org/Vol-4322/paper_9.pdf

1. Introduction

Knowledge graphs (KGs) have become a foundational technology for representing machine-actionable knowledge, enabling semantic search, automated reasoning, and intelligent knowledge discovery [1]. However, the utility of a KG is inherently bound to its quality [2]. Inconsistent vocabularies, incomplete metadata, inappropriate modeling decisions, and unstructured content reduce interoperability and limit the effectiveness of downstream applications [3]. Consequently, data management and curation have become a major research challenge in ensuring high-quality KGs [1].

Research Knowledge Graphs (RKGs) apply KG principles in the scholarly context by representing scientific knowledge as interconnected entities and relationships instead of unstructured documents [4]. One prominent and established RKG is the Open Research Knowledge Graph (ORKG) [5]. The ORKG enables researchers to collaboratively create structured representations of scientific knowledge. One of its central features are Comparisons, where contributors use shared entities and relationships to organize and compare knowledge from multiple publications to serve particular scholarly information needs and enable subsequent reuse by other users.

While this crowdsourcing model enables the rapid expansion of the ORKG, it presents significant data management and curation challenges. Contributors have broad modeling flexibility, deciding

DMKG 2026: 2nd International Workshop on Data Management for Knowledge Graphs co-located with ISWC 2026, Bari, Italy
lena.john@tib.eu (L. John); auer@tib.eu (S. Auer); oliver.karras@tib.eu (O. Karras)
0009-0007-2097-9761 (L. John); 0000-0002-0698-2864 (S. Auer); 0000-0001-5336-6899 (O. Karras)
© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-11-09

on granularity of data, assignment of data types, and the semantic representation of concepts. Every Comparison directly influences the structure, consistency, and interoperability of the graph, making the quality of the ORKG highly dependent on contributors adhering to consistent modeling practices.

To support high-quality contributions, the ORKG Curation & Community Building (C&CB) team has developed an ORKG modeling guideline [6] comprising recommendations covering quality of reuse, content, modeling, and representation. While this guideline consolidates previously established best practices into a coherent textual framework, its textual nature means it cannot readily be assessed automatically or translated into objective, real-time feedback during contribution workflows. This gap motivates the need for a comprehensive quality framework for curators and contributors that operationalizes the guideline by translating recommendations into quantifiable metrics.

Existing research has investigated KG quality dimensions such as completeness, consistency, accuracy, provenance, and interoperability [2, 7, 8]. However, most approaches focus on conceptual quality models or general graph quality assessment. Less attention has been paid to systematically translating modeling recommendations into operational quality metrics that assess the fitness for use of graphs [9], support contributors during content creation, and enable continuous quality monitoring.

In this paper, we present a quality framework for ORKG Comparisons that bridges this gap. Starting from the ORKG modeling guideline, we derive a set of operational quality metrics. These metrics are used for assessing the quality of existing content and for driving continuous monitoring during RKG modeling. As a first step toward the envisioned Comparison score, we implement a preliminary subset of these metrics in the ORKG curation dashboard SciKGDash [10] and apply them to current ORKG Comparisons. Our baseline analysis reveals recurring quality issues, including empty cells, incorrectly assigned data types, and inconsistent predicate label formatting, thereby validating the practicality of our approach. The presented quality framework establishes the foundation for future real-time quality feedback in the ORKG frontend, enabling contributors to receive actionable guidance while creating and editing Comparisons. The main contributions of this work are as follows:

1. A quality framework that maps ORKG modeling recommendations to operational quality metrics.
2. A Comparison-specific operationalization of the quality framework and a baseline assessment of ORKG Comparisons of the current modeling practices.
3. An implementation of the feasible Comparison quality metrics in SciKGDash, enabling continuous quality monitoring and providing curators with insights into recurring modeling issues.

The paper is structured as follows: Section 2 introduces background knowledge. Section 3 reviews related work. Section 4 presents the quality framework and approach. Section 5 describes the ORKG modeling guideline grouped into quality categories. Section 6 presents the operational metrics. Section 7 details the implementation. Section 8 presents the empirical assessment of ORKG content. Section 9 discusses the implications for future quality-aware contribution workflows. Section 10 presents future work and concludes the paper.

2. Background

The Open Research Knowledge Graph (ORKG) represents scholarly knowledge in a structured and semantic form by modeling research contributions as interconnected entities and relationships rather than unstructured text [5, 11]. Scientific knowledge is represented as subject–predicate–object triples, where a research contribution is typically the subject, predicates capture salient aspects of the contribution, and objects represent corresponding resources or typed literals. These triples are interconnected to form a knowledge graph, allowing concepts and their relationships to be represented and reused across different contributions. We use the general term *properties* instead of predicates in the following.

A central feature of the ORKG is Comparisons, which aggregate structured information from multiple research contributions addressing a common research problem [12]. A Comparison can be understood as a structured state-of-the-art overview in which multiple papers are represented using a common set of properties and values, allowing their reported characteristics and findings to be compared systematically. While Comparisons may address specific research questions or user needs, the concepts

used to represent their content become part of the shared RKG and can be reused across other ORKG artifacts and by other contributors. To facilitate their creation, the ORKG supports reusable templates that define common structures for recurring contribution types. However, modeling scientific findings remains challenging, requiring contributors to select appropriate levels of granularity, data types, and semantically consistent representations.

The ORKG can furthermore be viewed as a crowdsourcing platform in which contributors collaboratively create structured scholarly knowledge. Neto et al. [13] classify such systems as *crowdcontent* platforms, where participants collectively produce heterogeneous knowledge objects. Daniel et al. [14] distinguish between quality characteristics, quality assessment techniques, and quality assurance mechanisms as complementary aspects of quality management in crowdsourcing systems. In the ORKG, this flexibility is essential for capturing the diversity of scientific knowledge, but it also introduces the risk of heterogeneous and inconsistent modeling decisions across contributors.

The usefulness of an RKG ultimately depends on the quality of its content. Quality is commonly understood as the fitness of a graph with respect for its intended use [8, 9]. In this work, we assess quality in terms of compliance with recommended practices for representing structured scholarly knowledge, rather than the scientific adequacy of a Comparison for a particular use case. The proposed metrics therefore primarily indicate *modeling guideline conformance*, supporting the interoperability, maintainability, and machine-actionability of ORKG content. This provides the basis for systematically monitoring the quality of existing and newly created ORKG content.

3. Related Work

KG quality has been widely studied, resulting in several frameworks that define quality dimensions and corresponding evaluation metrics. Early work on Linked Data quality by Zaveri et al. [15] established a comprehensive taxonomy of quality dimensions, while Paulheim [16] surveyed methods for KG refinement, including error detection, completion, and consistency improvement. More recently, Hogan et al. [8] complement quality dimensions with corresponding quality metrics, where the dimensions describe qualitative characteristics of a KG and the metrics provide quantitative means to assess them. They emphasize that quality should be evaluated with respect to a KG's *fitness for purpose*.

Wang et al. [17] provide an overview of KG quality assessment by identifying six core quality dimensions (accuracy, completeness, consistency, timeliness, trustworthiness, availability) and discussing corresponding metrics. Huaman et al. [2] synthesize twenty quality dimensions from the literature and argue that their importance varies across application scenarios, requiring users to assign weights according to their specific needs. Chen et al. [18] complement these two frameworks with an application-oriented evaluation framework that explicitly links quality dimensions and metrics to the requirements of knowledge-based applications, thereby assessing whether a KG is fit for its intended purpose.

Within the ORKG, Hussein et al. [19] introduced the Knowledge Graph Maturity Model (KGMM), which defines five maturity levels for assessing the quality and maturity of research Comparisons. While KGMM provides guidance for improving Comparison quality, its hierarchical assessment can lead to low overall maturity ratings despite satisfying criteria from higher levels, and several of its measures rely on subjective expert judgement. Building on KGMM, D'Souza et al. [12] evaluated ORKG Comparisons and concluded that quality should be assessed through independent criteria tailored to specific use cases rather than aggregated into a single score. They further identify the need for more objective quality measures and explicit evaluation of semantic modeling.

Existing approaches provide comprehensive quality dimensions, metrics, and maturity models for KGs. However, they either target general-purpose KGs or assess quality at a high level of abstraction. In contrast, our work focuses on the ORKG curation process by operationalizing modeling guidelines as objective, measurable quality indicators. Rather than assigning maturity levels, we assess individual modeling aspects independently for a quality score or report, emphasizing objective measures of modeling quality while balancing machine-actionability with human usability.

4. Quality Framework

As part of maintaining and improving the ORKG, the Curation & Community Building (C&CB) team regularly reviews community contributions and supports contributors in RKG modeling. This process has revealed recurring modeling issues, including duplicate properties, inconsistent labels, missing descriptions, inappropriate data type usage, and the frequent use of long unstructured literals [10, 12, 20], affecting the consistency, interoperability, and maintainability of the ORKG negatively. To systematically address these issues, we propose a quality framework that combines modeling recommendations and computable quality metrics for supporting quality assessment. The framework is centered around the ORKG modeling guideline [6] that consolidates established best practices for representing scholarly knowledge and provides the conceptual foundation for operational quality assessment.

Figure 1 illustrates the envisioned quality improvement workflow involving three primary RKG stakeholder roles [21]: curator, developer, and contributor. Curators define and continuously refine the modeling guideline and its associated computable quality metrics. These artifacts guide developers in implementing quality assessment functionality across the ORKG ecosystem, including SciKGDash and the planned Comparison quality score, inspired by the KGMM [19], in the ORKG frontend.

SciKGDash provides curators with a global overview of community-generated content through automated quality assessment, interactive visualizations, and quality reports, enabling the identification of recurring modeling issues and opportunities for improving both the modeling guideline and the ORKG interface. In parallel, the frontend Comparison score supports contributors during the modeling process through live, advisory feedback based on selected quality metrics, allowing them to iteratively improve their Comparisons before publication without imposing publishing or modeling restrictions.

The resulting framework combines multiple quality management strategies following Daniel et al. [14]: computation-based quality assessment through operational metrics, contributor self-assessment through the Comparison score, and curator-driven review through SciKGDash (*fingerprinting*). Published Comparisons, together with aggregated quality analyses, provide curators with insights that

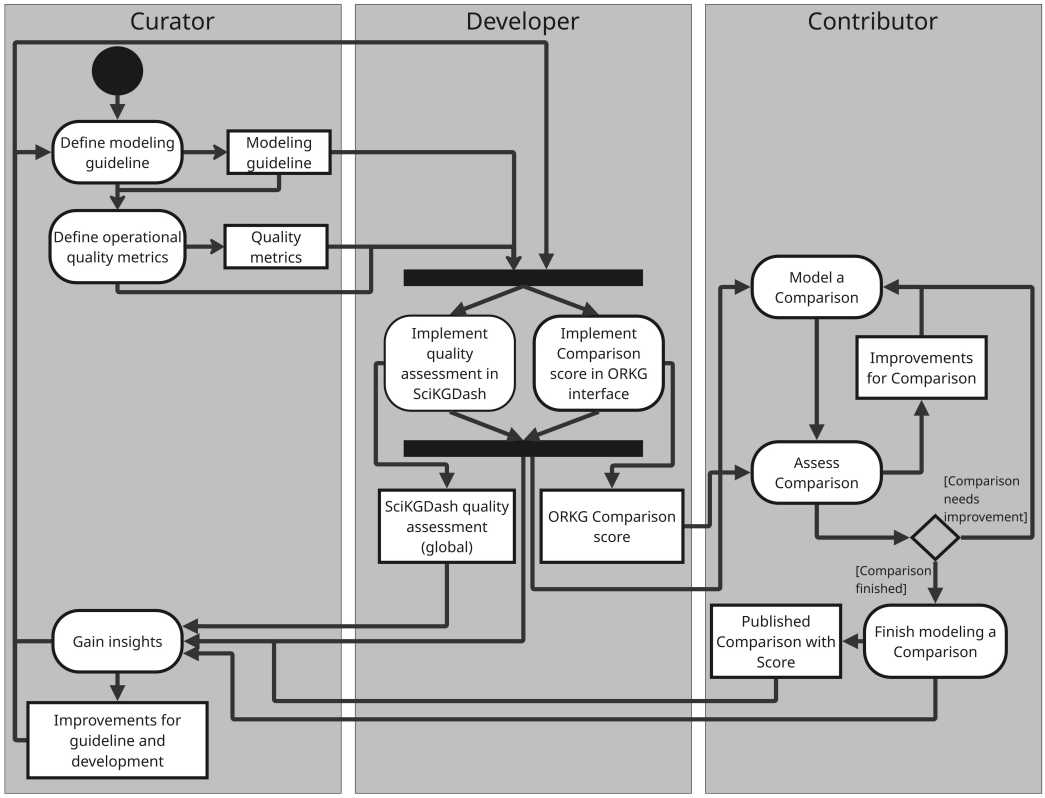


Figure 1: Activity diagram depicting the envisioned quality improvement workflow involving the three primary RKG stakeholder roles: curator, developer, and contributor. All three roles engage in continuous iterative activities.

inform further refinements of the guideline, metrics, and supporting tools. Curators, developers, and contributors continuously iterate through this quality improvement workflow, establishing an ongoing feedback loop for improving the quality and consistency of scientific knowledge in the ORKG.

In this work, we use the modeling guideline as the foundation for deriving operational quality metrics. As a first step towards the envisioned quality framework, we implement the metrics in SciKGDash, focusing on feasibility and Comparison-level metrics that can subsequently serve as the basis for a live Comparison quality score in the ORKG frontend. Beyond demonstrating the technical feasibility of the approach, the resulting implementation in the dashboard provides initial insights into the current quality of ORKG Comparisons and the overall ORKG content.

5. ORKG Modeling Guideline

The ORKG modeling guideline [6] translates established semantic modeling best practices into concrete recommendations for creating high-quality RKG contributions. Each recommendation follows a structured WHY–HOW–WHAT format, inspired by Sinek's Golden Circle [22] and has already been applied in two other guidelines [23, 24]. It first explains *why* adhering to the recommendation improves the overall quality of the ORKG, then specifies *how* contributors should model the respective aspect, and finally states *what* quality objective the recommendation aims to achieve. Where appropriate, recommendations are complemented by references to related recommendations and two examples.

The guideline was developed by consolidating existing ORKG best practices[1], educational material[2], and recurring modeling issues observed by the ORKG C&CB team. To provide a systematic perspective on quality, the recommendations were further aligned with established KG quality dimensions from the literature [2, 8, 17, 25, 26]. Since KG quality dimensions are not mutually exclusive and frequently overlap in practice, individual recommendations may contribute to multiple dimensions simultaneously.

The presented guideline represents the first iteration of an evolving quality framework for the ORKG. It intentionally focuses on the most common and broadly applicable modeling practices, particularly those relevant to ORKG Comparisons. Consequently, it should not be considered exhaustive and topics such as template-specific guidance or domain-dependent modeling conventions remain outside its current scope and will be incorporated in future iterations.

To organize the recommendations systematically, they are grouped into four quality categories: *Reuse Quality*, *Content Quality*, *Modeling Quality*, and *Representation Quality*. Together, these categories capture complementary aspects of scientific knowledge representation in ORKG. The individual recommendations describe concrete modeling practices that can subsequently be operationalized as measurable quality metrics (cf. Section 6). In the following, we introduce each quality category and present one exemplary recommendation from the modeling guideline [6].

Reuse Quality. Focuses on interoperability, reusability, and semantic consistency by encouraging contributors to reuse existing properties, resources, and external vocabularies, thereby reducing semantic duplication and increasing the interconnectedness of ORKG's content.
Content Quality. Focuses on completeness, accuracy, and trustworthiness by encouraging contributors to provide complete metadata and comparison content, ensuring that scholarly knowledge is represented correctly and sufficiently for its intended use.
Modeling Quality. Focuses on semantic accuracy and machine interpretability by promoting structured representations, appropriate data types, and hierarchical modeling patterns that support automated processing and reasoning.
Representation Quality. Focuses on consistency, understandability, and comparability by encouraging descriptive metadata, consistent naming conventions, and uniform modeling practices that improve the readability, maintainability, and discoverability of ORKG content.

In the following is an example for the category *Reuse Quality*:

[1]ORKG Best practices: https://orkg.org/help-center/article/42/Modeling_best_practices_for_resources_and_properties
[2]ORKG Academy: https://academy.orkg.org/concepts/semantic-modelling.html

R2: Reuse existing properties.
WHY: Creating duplicate properties fragments the knowledge graph. It prevents comparing resources accurately and results in maintenance efforts for cleaning up the knowledge graph. Reusing properties ensures your content instantly connects with the broader research ecosystem.
HOW:When a property is needed, always search the ORKG first to see if a matching or equivalent property already exists. Use the auto-complete suggestions, including the use of synonyms. Only create a new property if you are absolutely certain no existing property captures the meaning you need.
WHAT: A highly interconnected, standardized knowledge graph where different resources can be easily compared as they use the same vocabulary.

6. Operational Quality Metrics

The modeling guideline was, together with the C&CB team, systematically analyzed to identify measurable characteristics that reflect compliance with each recommendation. Since a single recommendation may address multiple aspects of modeling quality, one recommendation can give rise to several complementary quality metrics. Conversely, some recommendations are represented by a single metric when one measurement sufficiently captures the intended quality objective. The metrics were subsequently evaluated according to their technical feasibility. There is a total of 30 metrics.

The tables 1, 2, 3, and 4 present the metrics for each recommendation in the quality category. Implemented metrics are indicated by ☑ (23), planned metrics by ☐ (3) , and metrics that are currently infeasible by ↯(4), this is further discussed in the implementation Section 7. Depending on the scope, metrics can be assessed either globally across the entire ORKG, at the level of individual Comparisons, or at both levels. Global metrics provide insights into the overall state of ORKG content and recurring modeling practices, whereas Comparison-level metrics assess the quality of individual Comparisons and therefore form the foundation for the planned Comparison quality score.

The proposed metric framework is intentionally extensible. As the modeling guideline evolves, the technical conditions develop, and additional recommendations are introduced, new metrics can be incorporated without modifying the overall framework.

Table 1
Metrics (11) for measuring *Reuse Quality*

ID	Recommendation	Metric	Definition
R1	Create reusable properties and resources.	☑ R1.1 Property description	$\frac{\#\ properties\ without\ description}{\#\ all\ properties}$
		☑ R1.2 Class description coverage	$\frac{\#\ classes\ without\ description}{\#\ all\ classes}$
		☐ R1.3 Undocumented property usage rate	$\frac{\#\ used\ properties\ without\ description}{\#\ used\ properties\ in\ a\ Comparison}$
		☐ R1.4 Undocumented classes usage rate	$\frac{\#\ used\ classes\ without\ description}{\#\ used\ classes\ in\ a\ Comparison}$
R2	Reuse existing properties.	☑ R2.1 Duplicate property count	$\#\ duplicate\ properties$
		☑ R2.2 Property reuse rate	$\frac{\#\ properties\ used\ in\ multiple\ Comparisons}{\#\ all\ properties\ in\ a\ Comparison}$
		↯ R2.3 External property coverage	$\frac{\#\ properties\ linked\ to\ external\ vocabularies}{\#\ all\ properties\ in\ a\ Comparison}$
R3	Reuse existing resources.	☑ R3.1 Resource reuse rate	$\frac{\#\ reused\ resources}{\#\ all\ resources\ in\ a\ Comparison}$
		↯ R3.2 Duplicate resource count	$\#\ duplicate\ resources$
		☑ R3.3 External resource coverage	$\frac{\#\ resources\ linked\ to\ external\ vocabularies}{\#\ all\ resources\ in\ a\ Comparison}$
R4	Prefer resources over literals.	☑ R4.1 Resource-literal ratio	$\frac{\#\ resources}{\#\ all\ cells\ in\ a\ Comparison}$

Table 2
Metrics (3) for measuring *Content Quality*

ID	Recommendation	Metric	Definition
R5	Provide complete and correct metadata for a paper.	☑ R5.1 Metadata Completeness	$\frac{\#\ present\ metadata\ fields}{\#\ expected\ metadata\ fields\ in\ a\ paper}$
R6	Complete data entry in Comparisons.	☑ R6.1 Cell coverage	$\frac{\#\ filled\ cells}{\#\ all\ cells\ in\ a\ Comparison}$
		☑ R6.2 Explicit missing value rate	$\#$ "$Not\ reported$" $or\ similar\ entries$

Table 3
Metrics (5) for measuring *Modeling Quality*

ID	Recommendation	Metric	Definition
R7	Prefer machine-actionable, structured descriptions instead of using full sentences, long literals, long unstructured text.	☑ R7.1 Long literal count	$avg.\ literal\ length\ in\ a\ Comparison$
		☑ R7.2 Sentence-like literals	$\#\ literals\ containing\ sentences$
		☑ R7.3 Sentence-like resources	$\#\ resources\ containing\ sentences$
R8	Use appropriate data types.	☑ R8.1 Data type compliance	$\frac{\#\ correctly\ typed\ literals}{\#\ all\ literals\ in\ a\ Comparison}$
R9	Use structured representations to describe scientific content.	☑ R9.1 Hierarchy depth	$hierarchy\ depth\ of\ a\ Comparison$

Table 4
Metrics (11) for measuring *Representation Quality*

ID	Recommendation	Metric	Definition
R10	Write a good label.	☑ R10.1 Singular label compliance	$\frac{\#\ singular\ property\ labels}{\#\ all\ property\ labels}$
		☑ R10.2 Separator compliance	$\frac{\#\ property\ labels\ containing\ underscores}{\#\ all\ property\ labels}$
		☑ R10.3 Prefix compliance	$\frac{\#\ labels\ not\ starting\ with\ filler\ words}{\#\ all\ property\ labels}$
		☑ R10.4 Capitalization compliance	$\frac{\#\ labels\ following\ capitalization\ recommendations}{\#\ all\ property\ labels}$
		☑ R10.5 Label conformance score	$\frac{\#\ labels\ satisfying\ all\ rules}{\#\ all\ property\ labels}$
R11	Write a good description for properties.	☑ R11.1 Property description length	$avg.\ description\ length\ of\ properties$
		☐ R11.2 Class description length	$avg.\ description\ length\ of\ classes$
R12	Provide a description for a Comparison.	☑ R12.1 Description availability	$\frac{\#\ Comparisons\ with\ description}{\#\ all\ Comparisons}$
		☑ R12.2 Description length	$avg.\ description\ length\ of\ Comparisons$
R13	Use the same instances of entities and relationships across a Comparison.	↯ R13.1 Entity representation consistency	$\frac{\#\ repeated\ entities}{\#\ all\ entities\ in\ a\ Comparison}$
		↯ R13.2 Property representation consistency	$\frac{\#\ repeated\ properties}{\#\ all\ properties\ in\ a\ Comparison}$

7. Implementation

We implemented a large subset of the proposed metrics (metrics tagged with ☑, total of 23/30) in SciKGDash. While SciKGDash already provides graph-wide quality analyses [10, 20], such as description coverage (R1.1, R1.2), duplicate predicate detection (R2.1), and paper-related statistics (R5.1), this work extends the dashboard with a dedicated focus on ORKG Comparisons, which constitute the primary artifact targeted by the envisioned contributor-facing Comparison score.

The implemented analysis pipeline consists of four stages: (i) retrieval of Comparison contents, (ii) metric computation, (iii) aggregation of results across all analyzed Comparisons, and (iv) presentation within SciKGDash. To identify Comparisons for analysis, SciKGDash queries the ORKG triplestore via SPARQL to retrieve the identifiers and labels of all published Comparisons. It then retrieves the content of each Comparison through the ORKG contents endpoint (`/api/comparisons/id/contents`), which provides JSON representations of contributions, properties, literals, and resources.

For each recommendation, we first identify an observable property of the graph that reflects the intended modeling practice and then define a deterministic measurement procedure over the available ORKG data. Depending on the metric, this procedure operates on the structure of a Comparison, on the values contained in its cells, or on relationships between resources and properties reused across

the graph. Some constraints could alternatively be expressed using SHACL or ShEx, but the current ORKG guideline is not formalized as shape specifications. We distinguish three levels: *structural analyses* inspect the graph structure directly (e.g., hierarchy depth and cell coverage), *content analyses* evaluate values against deterministic rules (e.g., data type compliance, literal characteristics, and label conformance), and superficial *semantic analyses* require comparing content with their surrounding graph structure (e.g., resource reuse and the balance between resources and literals). Following computation, results are aggregated across all analyzed Comparisons to derive graph-wide observations and identify recurring modeling patterns. Table 5 summarizes the currently implemented metrics together with their corresponding SciKGDash routes. While several routes provide graph-wide analyses, the Comparisons route serves as the primary entry point for Comparison-level quality assessment and will therefore be the focus of the subsequent analysis.

Table 5
Implemented metrics and corresponding SciKGDash routes, additionally mentioning infeasible metrics.

SciKGDash Route	Implemented Metrics
/overview	R1.1, R1.2
/predicates	R2.1
/papers	R5.1
/comparisons	R2.2, R3.1, R3.3, R4.1, R6.1, R6.2, R7.1, R7.2, R7.3, R8.1, R9.1, R10.1, R10.2, R10.3, R10.4, R10.5, R11.1, R12.1, R12.2
Not possible	R2.3, R3.2, R13.1, R13.2

SciKGDash exposes the aggregated results through three complementary views. The *Charts* tab provides a histogram for each metric. The *List* tab presents a sortable and paginated overview of all Comparisons together with their computed metric values. Finally, the *Analysis Report* tab summarizes aggregated findings and provides more detailed analyses.

Metric selection (cf. Section 6) was guided by technical feasibility within the current ORKG infrastructure and the availability of the required data. While most implemented metrics can be computed directly from the Comparison contents endpoint, others remain impractical or infeasible. In particular, duplicate resource detection (R3.2) requires comparing potentially nested resource structures and assessing their semantic similarity, making exhaustive automatic detection computationally expensive. This is reflected in the existing duplicate predicate detection (R2.1) feature in SciKGDash, which uses human-in-the-loop assessment [20]. Entity representation consistency (R13.1) is more localized to an individual Comparison, while property representation consistency (R13.2) similarly requires semantic analysis and may require human judgment. External property coverage (R2.3) is currently not feasible because the ORKG backend does not track external property mappings; inferring them from prefixes alone would be unreliable and would require additional backend support. The planned features (R1.3, R1.4, R11.2) require additional endpoint calls and remain open for further optimization.

Several envisioned metrics would additionally benefit from provenance information and historical graph states. The ongoing migration of the ORKG backend toward event modeling and event sourcing [27] is expected to enable a new set of quality metrics based on graph evolution and contributor behavior, including analyses of modification histories, entity lifecycles, and the temporal development of Comparisons. Furthermore, future versions of SciKGDash [28] are expected to leverage the existing middleware database to persist intermediate results, thereby reducing API utilization and enabling more computationally demanding analyses.

8. Analysis

For the analysis, we will go through each of the guideline categories: *Reuse Quality*, *Content Quality*, *Modeling Quality*, and *Representation Quality*, and summarize the findings for the corresponding metrics. Unless stated otherwise, reported values correspond to Comparison-level metrics aggregated across all Comparisons. Graph-level metrics provide additional context regarding the overall state of ORKG

content. The metrics capture different aspects of Comparison quality. In general, higher values indicate better reuse, completeness, and compliance, while lower values indicate less redundancy, missing content, and violations, while metrics without a clear direction are interpreted descriptively.

We analyzed 1,789 published ORKG Comparisons after removing previous outdated versions, which would otherwise increase the dataset to 2,278 Comparisons. The preliminary Comparison score is calculated based on metrics that expose a percentage value, thereby not being of high expressivity. The score achieved an average of 55.6% (median: 57%), ranging from 17.4% to 100%.

RQ 1 (Reuse Quality): To what extent do ORKG Comparisons promote interoperability and reuse through the consistent use of entities and relationships?

Globally, approximately 77% of properties (R1.1) and 65% of classes (R1.2) lack descriptions, highlighting a substantial documentation gap. In addition, previous work [20] reported approximately 30% predicate redundancy (R2.1) in the ORKG, suggesting that duplicate properties remain a recurring challenge. Property reuse in Comparisons is comparatively high, with an average reuse rate of 79.9% (median: 93.3%) and 854 Comparisons exclusively reusing existing properties (R2.2). Resource reuse is considerably lower, averaging 37.9% (median: 28.6%), while 426 Comparisons do not reuse any resources at all (R3.1). This observation likely reflects the heterogeneous nature of ORKG resources, which are often accompanied by additional statements and nested structures that hinder straightforward reuse and duplicate detection. Resources linking to external vocabularies remain sparse, with 1,351 Comparisons lacking external resource references entirely (R3.3). Finally, the average resource-to-literal ratio of 51% and median of 53.5% (R4.1) indicate a balanced use of resources and literals.

RQ 2 (Content Quality): To what extent do ORKG Comparisons provide complete scientific content?

Paper metadata is generally well maintained (R5.1), with titles present for 99.99% of papers, author information available for 97.9%, and DOIs provided for 71.4% of publications. Most Comparisons are largely complete, although missing values remain common. On average, 9.9% of Comparison cells are empty (R6.1), while 758 Comparisons contain no empty cells. The median Comparison contains only 2 empty cells (2.2%). Explicitly modeled missing values are comparatively rare, with only 238 Comparisons containing placeholders (Not reported, Value not reported in paper, Empty) and a total of 2,990 placeholder values across the dataset (R6.2). However, these values are represented both as literals (768 occurrences) and resources (2,222 occurrences), and the analysis is likely to underestimate their true prevalence due to additional variations in phrasing. This observation highlights the absence of a canonical representation for explicitly missing values within the ORKG.

RQ 3 (Modeling Quality): To what extent do ORKG Comparisons employ semantically accurate and machine-interpretable modeling practices?

Literals are generally concise, with an average length of 26 characters (median: 10), although some reach up to 723 characters (R7.1). In addition, we identified 5,935 sentence-like literals across 526 Comparisons (R7.2) and 819 sentence-like resources across 204 Comparisons (R7.3), indicating that contributors occasionally prefer unstructured representations over machine-actionable entities.

ORKG currently supports the XML Schema built-in data types. Our analysis is limited to syntactically, not semantically, assessable XML data types, using `xsd:string` as the fallback type. Data type compliance (R8.1) is generally high, averaging 83.3% (median: 100%). A total of 1,043 Comparisons exhibit full data type compliance. Among those, 150 contain no literals at all, meaning that full compliance in these cases reflects the absence of values requiring data type validation rather than correct typing of literals. Furthermore, 73 Comparisons achieve no compliance at all. Across all 168,863 analyzed literals, we identified 27,513 data type mismatches, corresponding to an overall error rate of 16.3%. The data type compliance matrix (see Figure 2) further reveals that most violations stem from

values represented as `xsd:string` despite more specific data types being applicable, with integers incorrectly modeled as strings as the highest total error (9,473 occurrences). While some mismatches are attributable to difficulties in identifying the correct semantic XML data type (e.g., `xsd:gDay` interpreted as `xsd:string`), the results indicate that contributors frequently rely on default string representations.

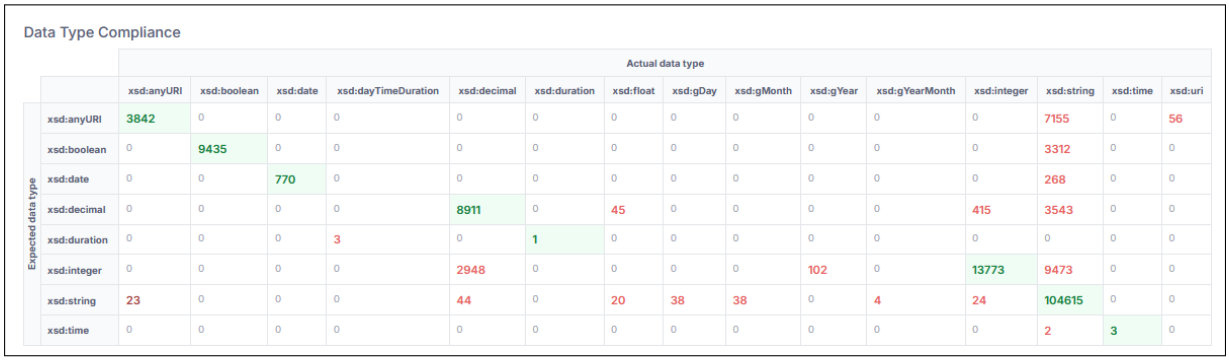


Figure 2: Data type mismatch matrix comparing expected and given XML XSD data types. Most commonly, users type literals as the default string data type.

Finally, hierarchical modeling remains uncommon (R9.1), as 1,258 Comparisons exhibit a hierarchy depth of 0, while only 27 reach the maximum observed depth of 4. Overall, ORKG Comparisons largely adhere to data type recommendations but rarely exploit more advanced structured modeling patterns.

RQ 4 (Representation Quality): To what extent do ORKG Comparisons conform to established representation practices that support consistency, understandability, and comparability?

Representation quality remains mixed. Across 20,659 property labels, 11,966 exhibit at least one violation of the labeling recommendations, resulting in an overall quality score of 42.1% (R10.5). The average label quality score was 43.5% (median: 40%). The most frequent issues are labels starting with uppercase characters (35.7%) (R10.4) and the use of plural forms (19.4%) (R10.1), followed by verb prefixes such as "has", "is", or "uses" (10.2%) (R10.3). Violations related to separators are comparatively uncommon, with only 2.5% (R10.2) of labels containing underscores and 1.1% using camelCase.

Property descriptions exhibit an average length of 89.9 characters (median: 63) (R11.1), i.e. one or two short sentences. Of the 11,160 properties that provide a description, 168 contain ≤ 10 characters, using a heuristic threshold for unusually short descriptions. While Comparison descriptions are mandatory for publishing in ORKG, we nevertheless identified 166 Comparisons without descriptions, corresponding to an overall description coverage of 90.7% (R12.1). Manual inspection suggests that these instances primarily originate from legacy or testing data present in the production environment. Comparison descriptions exhibit an average length of 343 characters (median: 211) (R12.2), with observed lengths ranging from 0 to 5,716 characters, indicating that contributors generally provide concise textual summaries while a small number of Comparisons contain substantially more extensive descriptions.

To complement the aggregated statistics, SciKGDash enables the identification of individual Comparisons exhibiting notable modeling characteristics. For example, the Comparison with the lowest overall score (17.4%) exhibits multiple quality issues, including non-conforming property labels, missing property descriptions, frequent use of string literals where typed values or resources would be more appropriate, and 46.9% empty cells. Moreover, semantic structures such as time spans are modeled as strings rather than structured entities, illustrating possible extensions of the current metric set. Conversely, the Comparison with the highest score (74.5%) among those with the maximum hierarchy depth of 4 demonstrates desirable modeling practices, including complete property reuse, present property descriptions (90%), and a balanced use of resources and literals (52.3% resources). However, manual inspection reveals that some deeply nested structures redundantly represent the same concept and could be modeled more directly. These examples illustrate both the usefulness of SciKGDash and limitations of automated metrics, highlighting the continued need for qualitative and human inspection

to assess semantic modeling quality. Furthermore, these observations suggest a gap between existing best practices and their application in practice.

9. Discussion

The presented metrics provide a first quantitative view of how contributors model scientific knowledge in ORKG Comparisons. However, the analysis shows that several quality aspects, particularly semantic correctness and scientific adequacy, cannot be captured by structural metrics alone.

For *reuse quality*, property reuse is comparatively high, although previous work reported approximately 30% property redundancy across the ORKG [20], indicating that reuse does not necessarily imply the use of canonical properties. Resource reuse and external resource linking remain considerably lower, likely reflecting the heterogeneous and often nested nature of ORKG resources, which complicates straightforward reuse and duplicate detection. The absence of a duplicate resource metric currently prevents assessing semantic redundancy at the resource level. Reuse quality can therefore only be assessed partially and should be complemented by graph-wide redundancy measures in future analyses.

Content quality metrics indicate that Comparisons are generally structurally complete, with well-maintained paper metadata and most cells populated. However, the current metrics cannot assess semantic completeness, i.e. whether modeled values are correct with respect to the original publication, whether relevant information is missing, or whether a Comparison is scientifically meaningful and comparable. These aspects require human validation. The inconsistent representation of missing values exposes a modeling limitation of ORKG itself. Since template fields often expect entities of a particular type or class, contributors cannot always provide a canonical "Not reported" value, motivating the need for a type-independent representation of explicitly missing information.

Regarding *modeling quality*, data type compliance is generally high, but the analysis is limited to syntactic validation of XML data types. Semantic validation remains an open challenge, e.g., an integer literal may be syntactically correct but semantically represent a calendar day or a duration depending on the modeled context. Hierarchical structures are rarely used, suggesting that contributors either do not exploit or are insufficiently supported in using ORKG's semantic modeling capabilities. Template- or research field-specific analyses could provide additional insights into current modeling practices.

Representation quality reveals that several modeling recommendations are not consistently applied by contributors. Inconsistent property labels, numbered properties (e.g., 1st Dataset Accuracy), and descriptions containing actual values rather than explanatory text (e.g., Location) suggest that recommendations are not sufficiently visible during the contribution process. Several issues could be detected automatically and addressed through real-time recommendations in the ORKG frontend. Rather than penalizing contributors, such metrics may therefore be more useful as warnings or actionable feedback.

More generally, the analysis demonstrates that designing and interpreting a single aggregated Comparison score is challenging. High scores do not necessarily correspond to semantically meaningful or scientifically useful Comparisons, whereas lower scores may merely reflect a small number of guideline violations despite otherwise well-modeled content. Instead of a single score, category-specific progress indicators for reuse, content, modeling, and representation quality may provide more transparent and actionable feedback. Quantifying the impact of individual dimensions would require a hierarchical quality model that relates and weighs metrics with respect to overall quality [2, 8] and the intended use of a Comparison. The relevance of a quality dimension also depends on the intended use of a Comparison. For instance, inconsistent property labels may primarily hinder discovery and reuse, whereas missing hierarchical structures may reduce the semantic interpretability of the represented knowledge for humans and machines. In the current work, fitness for use is therefore operationalized more narrowly as compliance with recommended practices for representing structured scholarly knowledge, rather than the scientific adequacy of the represented content.

The envisioned Comparison score can combine these complementary aspects: curators can define general expectations for how scientific knowledge should be represented, in form of modeling recommendations, while contributors provide the domain-specific knowledge required to determine what

should be represented and whether the resulting Comparison is scientifically meaningful. The score can thus serve as a collaborative assessment of fitness for use, combining automated assessment of modeling practices with domain expertise. Future work can then investigate questions as such: How do individual quality dimensions influence the fitness for use of a Comparison across different use cases?

Although the concrete metrics are derived from ORKG's modeling guideline, their underlying quality dimensions are applicable to KGs more generally. Developing a robust Comparison quality model will require aligning these application-specific metrics with established KG quality models [8, Chapter 2.4.2], defining hierarchical quality (sub-)characteristics, an enhanced metric set, and corresponding evaluation procedures [29].

10. Future Work and Conclusion

Future Work. Future work will focus on advancing into a hierarchical quality model for ORKG Comparisons by aligning the application-specific metrics with established KG quality models, such as Hogan et al. [8]. Quality (sub-)characteristics will be operationalized through measurable metrics that form the foundation of the envisioned Comparison score. Rather than merely assigning a score, the frontend implementation will provide contributors with live, actionable feedback during the modeling process, directly communicating the modeling guideline and supporting learning-by-doing.

Furthermore, the guideline itself will be refined and extended with more detailed recommendations, including advanced ORKG features such as templates or classes. Additional curator support mechanisms are required to improve Comparison quality after creation. In this regard, SciKGDash will continue to support graph-wide quality monitoring and curation workflows for curators. A promising direction is to extend SciKGDash to elevate contributors to curators of their own data by providing personalized quality assessments of their ORKG content together with data cleaning and improvement workflows. This could help contributors better understand the consequences of their modeling decisions and increase ownership of the created content. Finally, several metrics remain technically infeasible and require further collaboration with the ORKG backend team. In particular, duplicate resource detection and semantic consistency analyses require graph traversal and more computationally demanding analyses, with the latter constituting a separate research direction. Future versions of SciKGDash are intended to leverage middleware-supported intermediate results to enable such analyses while improving scalability.

Conclusion. In this paper, we presented a quality framework that derives measurable quality metrics from the ORKG modeling guideline [6] and established the foundation for a future Comparison score. The first implementation in SciKGDash provides quantitative insights into current ORKG Comparison modeling practices and recurring quality issues. Our findings demonstrate that structural quality can be assessed automatically to a certain extent, while semantic correctness and scientific adequacy remain challenges that require further methodological and human-centered quality assessment approaches.

Declaration on Generative AI

During the preparation of this work, the authors used Gemini 3.5 Flash and GPT 5.5 to: Abstract drafting, Drafting content, Paraphrase and reword, Grammar and spelling check. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the paper's content.

References

[1] D. Fensel, et al., Knowledge Graphs-Methodology, Tools and Selected Use Cases, Springer, 2020.

[2] E. Huaman, D. Fensel, Knowledge Graph Curation: A Practical Framework, in: 10th International Joint Conference on Knowledge Graphs, 2022.

[3] E. Huaman, et al., Duplication Detection in Knowledge Graphs: Literature and Tools, ArXiv (2020).

[4] O. Karras, et al., Divide and Conquer the EmpiRE: A Community-Maintainable Knowledge Graph of Empirical Research in Requirements Engineering, in: 17th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM), IEEE, 2023.
[5] S. Auer, et al., Open Research Knowledge Graph: A Large-Scale Neuro-Symbolic Knowledge Organization System, in: Handbook on Neurosymbolic AI and Knowledge Graphs, IOS Press, 2025.
[6] O. Karras, et al., Technical Report: Curation Modeling Guideline for the Open Research Knowledge Graph, TIB Hannover : Repositorium für Naturwissenschaften und Technik (2026).
[7] B. Xue, et al., Knowledge Graph Quality Management: A Comprehensive Survey, IEEE Trans. on Knowledge and Data Engineering (2022).
[8] A. Hogan, et al., Knowledge Graphs, Springer, 2022.
[9] D. M. Strong, Y. W. Lee, R. Y. Wang, Data quality in context, Communications of the ACM (1997).
[10] L. John, et al., SciKGDash: The Scientific Knowledge Graph Dashboard for Supporting Knowledge Curation, in: ACM/IEEE Joint Conference on Digital Libraries, IEEE, 2025.
[11] M. Stocker, et al., SKG4EOSC - Scholarly Knowledge Graphs for EOSC: Establishing a Backbone of Knowledge Graphs for FAIR Scholarly Information in EOSC, RIO 8 (2022).
[12] J. D'Souza, et al., Quality Assessment of Research Comparisons in the Open Research Knowledge Graph, JLIS.It 15 (2024).
[13] F. R. A. Neto, C. A. Santos, Understanding crowdsourcing projects: A systematic review of tendencies, workflow, and quality management, Information Processing & Management (2018).
[14] F. Daniel, et al., Quality Control in Crowdsourcing: A Survey of Quality Attributes, Assessment Techniques, and Assurance Actions, ACM Comput. Surv. (2018).
[15] A. Zaveri, et al., Quality assessment for Linked Data: A Survey: A systematic literature review and conceptual framework, Semantic Web 7 (2015).
[16] H. Paulheim, P. Cimiano, Knowledge graph refinement: A survey of approaches and evaluation methods, Semantic Web (2016).
[17] X. Wang, et al., Knowledge Graph Quality Control: A survey, Fundamental Research (2021).
[18] H. Chen, et al., A Practical Framework for Evaluating the Quality of Knowledge Graph, Knowledge Graph and Semantic Computing: Knowledge Computing and Language Understanding (2019).
[19] H. Hussein, et al., KGMM - A Maturity Model for Scholarly Knowledge Graphs Based on Intertwined Human-Machine Collaboration, ICADL (2022).
[20] L. John, et al., Addressing Predicate Redundancy in Research Knowledge Graphs: Duplicate Detection, Resolution, and Prevention, in: Linking Theory and Practice of Digital Libraries (TPDL 2026), Springer, 2026.
[21] H. Li, et al., Knowledge Graphs in Practice: Characterizing their Users, Challenges, and Visualization Opportunities, IEEE Trans. on Visualization and Computer Graphics (2024).
[22] S. Sinek, Start with Why: How Great Leaders Inspire Everyone to Take Action, Penguin, 2009.
[23] O. Karras, The Ethos of the PEERfect REVIEWer: Scientific Care and Collegial Welfare, in: Joint Proceedings of REFSQ-2026 Workshops, Doctoral Symposium, Posters & Tools Track, and Education and Training Track. Co-located with REFSQ 2026, CEUR, 2026.
[24] O. Karras, K. Schneider, An Interdisciplinary Guideline for the Production of Videos and Vision Videos by Software Professionals, arXiv preprint arXiv:2001.06675 (2021).
[25] C. Batini, M. Scannapieco, Data and Information Quality, Springer, 2016.
[26] C. Cichy, S. Rass, An overview of data quality frameworks, IEEE Access (2019).
[27] M. Dilger, Understanding eventsourcing, Independently published, 2024.
[28] L. John, S. Auer, O. Karras, SciKGDash: The Human-in-the-Loop Curation Dashboard for the Open Research Knowledge Graph, in: Proceedings of the 25th International Semantic Web Conference (ISWC 2026), Posters & Demos Track, 2026.
[29] D. L. Moody, Theoretical and practical issues in evaluating the quality of conceptual models: Current state and future directions, Data & Knowledge Engineering 55 (2005).