Cost-Effective Quality Evaluation for Evolving Knowledge Graphs

STEFANO MARCHESIN, University of Padua, Italy
GIANMARIA SILVELLO, University of Padua, Italy
OMAR ALONSO*, Amazon, USA

ACM Reference Format:
Stefano Marchesin, Gianmaria Silvello, and Omar Alonso. 2026. Cost-Effective Quality Evaluation for Evolving Knowledge Graphs. *ACM J. Data Inform. Quality* 1, 1 (July 2026), 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Knowledge Graph Quality Evaluation

Most research on Knowledge Graph (KG) quality adopts a *source-centric view* that prioritizes identifying reliable data sources [5, 7, 8], documenting ingestion and transformation pipelines [25, 32], and specifying configuration artifacts that capture which sources contribute to which predicates and entity types [9, 15]. As a result, a large body of work has examined source reliability metrics [12, 34], pipeline governance [33], and best practices for selecting authoritative inputs [10, 35]. These ideas governed the design of widely used KGs such as DBpedia [2], YAGO [29], and Wikidata [31], which incorporate references and provenance from external sources to support facts. Following these accounts, we use *quality evaluation* to denote the systematic, evidence-bearing measurement of fitness-for-use properties of a KG (e.g., accuracy, consistency, completeness, timeliness [35]) via a combination of automated validation (e.g., SHACL constraint checking), manual verification (e.g., sampling-based accuracy estimation [13]), and provenance-aware audits [10, 33].

This view is appealing: by prioritizing trusted inputs and transparent pipelines, data creators and curators use source credibility as a proxy for KG reliability. Foundational work on KG trustworthiness [6, 8] and large-scale KG construction [4, 32] identifies source credibility, curation workflows, and systematic ingestion as prerequisites for quality [21, 29]. Industrial-scale KG systems likewise emphasize careful source selection and rigorous pipeline governance [16, 17, 23].

At the same time, recent work has started to highlight how the source-centric view might be limited for modern KGs in dynamic, evolving domains [22, 27]. In practice, many widely used KGs continuously evolve as facts change, new evidence emerges, and contributions accumulate. Quality challenges in these settings extend beyond source selection: systems must also explain how facts evolve over time, determine when previously verified facts require re-evaluation, and provide evidence that downstream applications can interpret and audit. Hence, we identify data-quality

Preprint (authors version)

*Work does not relate to the author's position at Amazon.

Authors' Contact Information: Stefano Marchesin, stefano.marchesin@unipd.it, University of Padua, Padua, Italy; Gianmaria Silvello, gianmaria.silvello@unipd.it, University of Padua, Padua, Italy; Omar Alonso, omralon@amazon.com, Amazon, Santa Clara, California, USA.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2026 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM 1936-1963/2026/7-ART
https://doi.org/10.1145/nnnnnnn.nnnnnnn

challenges that arise when KGs evolve continuously, with particular emphasis on temporal validity, provenance, and cost-aware re-evaluation.

To this end, let us consider representative scenarios: (E1) a gene is reclassified from tumor suppressor to oncogene as scientific understanding evolves; (E2) a professional athlete transfers between teams; (E3) a corporate acquisition takes place, triggering a cascade of dependent updates (subsidiary relations, executive roles, stock identifiers). Unlike standard correction cases, where an assertion is revised because it was simply the result of a human or automated error, E1–E3 reflect actual changes in the real world. As a result, correctness depends not only on source reliability, but also on the temporal dimension: when an assertion entered the KG, when it held, and when it was modified or retracted. Evolving KGs must therefore satisfy three key requirements.

R1: Temporal provenance and history (what the KG records). The KG must store, alongside each assertion, its temporal scope, the evidence supporting it, and a retrievable history of prior states. In E1, R1 mandates that `gene-X is-a tumor-suppressor` be retrievable together with its 2019 evidence and validity interval, distinct from the 2023 reclassification as oncogene; this would enable retrospective and longitudinal queries (e.g., "What function was attributed to gene X in 2019 compared with 2023?"). R1 is a *representation* requirement on the data and metadata.

R2: Context-dependent correctness (how correctness is assessed). Whether an assertion is correct must be evaluated *against the time and evidence to which it refers*, not against a single, atemporal ground truth. In E2, `plays-for(A, Team1)` is correct in 2022 and incorrect in 2024, and both judgments must be supported simultaneously by the evaluator given what R1 stored. R2 is a *judgment* requirement on the evaluation procedure, and it presupposes R1.

R3: Cost-aware quality management (how often, how thoroughly). Sustaining KG quality under continuous change requires budget-aware monitoring and updates. The evaluation process must decide *when* to recompute quality estimates, *which* assertions to re-test, and *how* to trade off computational cost (e.g., human annotation, snapshot storage) against expected quality gain [13]. In E3, R3 governs which of the corporate acquisition's dependent updates are re-verified eagerly and which are deferred. This involves choices such as keeping full historical snapshots vs. compressing high-velocity data (e.g., stock prices or game scores), or refreshing quality measures after every update vs. only when changes exceed a threshold.

The challenge we discuss in this paper is therefore not whether KGs can be updated, but how their quality can be assessed and maintained once updates become continuous, heterogeneous, and temporally consequential:

> *Which data-quality issues arise in a continuously evolving KG, and what evidence is required to detect, explain, and prioritize them under limited verification budgets?*

To make this challenge space explicit, we distinguish four recurring *dimensions* of quality assessment in evolving KGs. The first is the **data source dimension (D1)**: when sources change, disagree, or drift in coverage and reliability, source selection alone no longer guarantees the quality of the resulting KG (E1: a paper retracted in 2024 was the original evidence for the 2019 assertion). The second is the **reliability dimension (D2)**: validation cannot be reduced to a one-time release check, because previously accepted assertions may become incorrect, unsupported, or inconsistent *after* later updates (E2: checking for `plays-for(A, Team1)` only at ingestion misses the 2024 transfer). The third is the **temporal dimension (D3)**: many assertions are meaningful only relative to a validity interval, and quality depends on whether the KG records not only *what* is asserted, but also *when* it held and was observed (E1, E2, E3 all require validity intervals). The fourth is the **utility dimension (D4)**: with realistic budgets, not every assertion can be re-verified after each change, so systems must prioritize verification effort and assess the risk of deferring it (E3: thousands of dependent updates compete for a fixed annotation budget).

Preprint (authors version)

The four dimensions partition the *origin* of quality issues (source, reliability, temporal) from the *constraint* on the response (utility), and they encode the requirements: R1 records along source, temporal, and (partial) reliability; R2 reads them to issue judgments; R3 imposes the utility constraint on R1 and R2. The rest of the paper discusses the open challenges that this framing exposes.

2 Open Challenges

Realizing the three requirements along the four dimensions raises five interrelated challenges.

C1: Data and metadata management at scale. Explicit provenance, evidence, and temporal validity increase the volume of metadata and intermediate artifacts that must be stored, indexed, and kept up to date – potentially shifting cost from "verification" to "metadata freshness". This is not hypothetical: large, collaborative – and possibly interlinked – KGs are updated at high rates and accumulate massive histories, and practical systems already expose the tension between update throughput and metadata handling [27, 30]. At the same time, keeping multiple KG versions leads to well-known scalability issues in RDF archiving: naive version storage does not scale, and snapshot-delta strategies must balance ingestion time, storage size, and query performance [3, 11]. C1 maps to requirements R1, R3, and dimensions D1, D3.

C2: Heterogeneity of temporal semantics and update regimes. Time-aware KGs include at least two distinct objectives – modeling *validity periods* of facts and ensuring *traceability* of how and why facts changed – and the literature uses terms such as dynamic or temporal inconsistently [20]. Moreover, KGs span extremes in update cadence, from continuously synchronized extractions to stream-like settings. A practical quality discussion must therefore be explicit about which scenario is targeted (continuously updated releases, time-stamped facts, or both) and which assumptions are made about time granularity [18]. C2 maps to requirements R1, R2, and dimension D3.

C3: Representation and querying complexity. Attaching temporal validity and evidence at statement level introduces trade-offs in granularity, compression of unchanged facts, and indexing of time intervals. While work on temporal formalisms (e.g., temporal description logics [1]) and on temporal RDF models [14] provides conceptual foundations, engineering scalable representations without losing queryability remains non-trivial [3]. C3 maps to requirements R1, R2, and dimensions D2, D3.

C4: Limits of automated reliability testing. Constraint languages such as SHACL support systematic validation of RDF graphs, but they primarily capture structural and constraint-based checks; subtle semantic errors and temporal incoherencies may escape automated suites unless temporal constraints and domain semantics are made explicit and testable. At the same time, KGs may also evolve at the schema level. Designing "release tests" that combine constraint validation with temporal coherence across both instance and schema evolution remains largely open [27]. C4 maps to requirement R2 and dimension D2.

C5: Change-rate classification and adaptation. Operationally, deciding *what to re-verify when* depends on assumptions about how facts evolve; yet in real domains, change dynamics may shift, and misclassification can produce either staleness (under-verification) or wasted effort (over-verification). Detecting and correcting such misclassifications requires feedback signals from both history and validation outcomes, under realistic storage and ingestion constraints [26]. C5 maps to requirement R3 and dimension D4.

Overall, KG quality in evolving settings is an ongoing operational problem rather than a one-time engineering step. As KGs become infrastructure for applications that require time-sensitive and evidence-grounded reasoning [19, 28], and the interplay between KGs and LLMs gains prominence [24], the need grows for quality perspectives that jointly consider provenance, temporal validity, reproducible validation, and cost constraints. The five challenges therefore outline a research agenda whose priorities follow directly from them: C1-C3 call for temporally indexed,

Preprint (Authors version)

version-aware RDF stores that expose statement-level provenance with limited maintenance overhead [3, 11]; C4 calls for constraint frameworks extending SHACL with temporal coherence and domain-specific predicates, complemented by budget-aware, human-in-the-loop quality estimation with statistical guarantees on evolving releases [13]; C5 calls for change-rate models learned from validation history and edit logs [26, 30] that trigger targeted re-verification within explicit budgets. Treating these directions jointly, rather than as separate solutions, is the framing we envision: a quality discipline that records what is needed to judge (R1), evaluates judgments against the appropriate context (R2), and rations verification under realistic budgets (R3), composing across the four dimensions of quality assessment.

References

[1] A. Artale and E. Franconi. 2000. A survey of temporal extensions of description logics. *Ann. Math. Artif. Intell.* 30, 1-4 (2000), 171–210. doi:10.1023/A:1016636131405

[2] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. G. Ives. 2007. DBpedia: A Nucleus for a Web of Open Data. In *The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, Busan, Korea, November 11-15, 2007 (LNCS, Vol. 4825)*. Springer, 722–735. doi:10.1007/978-3-540-76298-0_52

[3] A. Cerdeira-Pena, G. de Bernardo, A. Fariña, J. D. Fernández, and M. A. Martínez-Prieto. 2024. Compressed and queryable self-indexes for RDF archives. *Knowl. Inf. Syst.* 66, 1 (2024), 381–417. doi:10.1007/S10115-023-01967-7

[4] O. Deshpande, D. S. Lamba, M. Tourn, S. Das, S. Subramaniam, A. Rajaraman, V. Harinarayan, and A. Doan. 2013. Building, maintaining, and using knowledge bases: a report from the trenches. In *Proc. of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2013, New York, NY, USA, June 22-27, 2013*. ACM, 1209–1220. doi:10.1145/2463676.2465297

[5] X. L. Dong, L. Berti-Équille, and D. Srivastava. 2009. Integrating Conflicting Data: The Role of Source Dependence. *Proc. VLDB Endow.* 2, 1 (2009), 550–561. doi:10.14778/1687627.1687690

[6] X. L. Dong, E. Gabrilovich, G. Heitz, W. Horn, N. Lao, K. Murphy, T. Strohmann, S. Sun, and W. Zhang. 2014. Knowledge vault: a web-scale approach to probabilistic knowledge fusion. In *Proc. of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24 - 27, 2014*. ACM, 601–610. doi:10.1145/2623330.2623623

[7] X. L. Dong, E. Gabrilovich, G. Heitz, W. Horn, K. Murphy, S. Sun, and W. Zhang. 2014. From Data Fusion to Knowledge Fusion. *Proc. VLDB Endow.* 7, 10 (2014), 881–892. doi:10.14778/2732951.2732962

[8] X. L. Dong, E. Gabrilovich, K. Murphy, V. Dang, W. Horn, C. Lugaresi, S. Sun, and W. Zhang. 2015. Knowledge-Based Trust: Estimating the Trustworthiness of Web Sources. *Proc. VLDB Endow.* 8, 9 (2015), 938–949. doi:10.14778/2777598.2777603

[9] L. Ehrlinger and W. Wöß. 2016. Towards a Definition of Knowledge Graphs. In *Joint Proc. of the Posters and Demos Track of the 12th International Conference on Semantic Systems - SEMANTiCS2016 and the 1st International Workshop on Semantic Change & Evolving Semantics (SuCCESS'16) co-located with the 12th International Conference on Semantic Systems (SEMANTiCS 2016), Leipzig, Germany, September 12-15, 2016 (CEUR Workshop Proceedings, Vol. 1695)*. CEUR-WS.org. https://ceur-ws.org/Vol-1695/paper4.pdf

[10] M. Färber, F. Bartscherer, C. Menne, and A. Rettinger. 2018. Linked data quality of DBpedia, Freebase, OpenCyc, Wikidata, and YAGO. *Semantic Web* 9, 1 (2018), 77–129. doi:10.3233/SW-170275

[11] J. D. Fernández, J. Umbrich, A. Polleres, and M. Knuth. 2019. Evaluating query and storage strategies for RDF archives. *Semantic Web* 10, 2 (2019), 247–291. doi:10.3233/SW-180309

[12] A. Galland, S. Abiteboul, A. Marian, and P. Senellart. 2010. Corroborating information from disagreeing views. In *Proc. of the Third International Conference on Web Search and Web Data Mining, WSDM 2010, New York, NY, USA, February 4-6, 2010*. ACM, 131–140. doi:10.1145/1718487.1718504

[13] J. Gao, X. Li, Y. E. Xu, B. Sisman, X. L. Dong, and J. Yang. 2019. Efficient Knowledge Graph Accuracy Evaluation. *Proc. VLDB Endow.* 12, 11 (2019), 1679–1691. doi:10.14778/3342263.3342642

[14] C. Gutierrez, C. A. Hurtado, and A. A. Vaisman. 2005. Temporal RDF. In *The Semantic Web: Research and Applications, Second European Semantic Web Conference, ESWC 2005, Heraklion, Crete, Greece, May 29 - June 1, 2005, Proc. (LNCS, Vol. 3532)*. Springer, 93–107. doi:10.1007/11431053_7

[15] A. Hogan, E. Blomqvist, M. Cochez, C. d'Amato, G. de Melo, C. Gutierrez, S. Kirrane, J. E. Labra Gayo, R. Navigli, S. Neumaier, A. C. Ngonga Ngomo, A. Polleres, S. M. Rashid, A. Rula, L. Schmelzeisen, J. Sequeda, S. Staab, and A. Zimmermann. 2021. *Knowledge Graphs*. Morgan & Claypool Publishers. doi:10.2200/S01125ED1V01Y202109DSK022

[16] I. F. Ilyas, J. Lacerda, Y. Li, U. F. Minhas, A. Mousavi, J. Pound, T. Rekatsinas, and C. Sumanth. 2023. Growing and Serving Large Open-domain Knowledge Graphs. In *Companion of the 2023 International Conference on Management of*

Preprint (authors version)

Data, SIGMOD/PODS 2023, Seattle, WA, USA, June 18-23, 2023. ACM, 253–259. doi:10.1145/3555041.3589672

[17] I. F. Ilyas, T. Rekatsinas, V. Konda, J. Pound, X. Qi, and M. A. Soliman. 2022. Saga: A Platform for Continuous Construction and Serving of Knowledge at Scale. In *SIGMOD '22: International Conference on Management of Data, Philadelphia, PA, USA, June 12 - 17, 2022.* ACM, 2259–2272. doi:10.1145/3514221.3526049

[18] X. Jiang, C. Xu, Y. Shen, X. Sun, L. Tang, S. Wang, Z. Chen, Y. Wang, and J. Guo. 2023. On the Evolution of Knowledge Graphs: A Survey and Perspective. *CoRR* abs/2310.04835 (2023). doi:10.48550/ARXIV.2310.04835

[19] S. M. Kazemi, R. Goel, K. Jain, I. Kobyzev, A. Sethi, P. Forsyth, and P. Poupart. 2020. Representation Learning for Dynamic Graphs: A Survey. *J. Mach. Learn. Res.* 21 (2020), 70:1–70:73. https://jmlr.org/papers/v21/19-447.html

[20] F. Krause, T. Weller, and H. Paulheim. 2022. On a Generalized Framework for Time-Aware Knowledge Graphs. In *Towards a Knowledge-Aware AI - SEMANTiCS 2022 - Proc. of the 18th International Conference on Semantic Systems, 13-15 September 2022, Vienna, Austria (Studies on the Semantic Web, Vol. 55).* IOS Press, 69–74. doi:10.3233/SSW220010

[21] S. Marchesin, L. Menotti, F. Giachelle, G. Silvello, and O. Alonso. 2023. Building a Large Gene Expression-Cancer Knowledge Base with Limited Human Annotations. *Database J. Biol. Databases Curation* 2023 (2023). doi:10.1093/DATABASE/BAAD061

[22] S. Mohammed, L. Ehrlinger, H. Harmouch, F. Naumann, and D. Srivastava. 2025. The Five Facets of Data Quality Assessment. *SIGMOD Rec.* 54, 2 (2025), 18–27. doi:10.1145/3749116.3749120

[23] N. F. Noy, Y. Gao, A. Jain, A. Narayanan, A. Patterson, and J. Taylor. 2019. Industry-scale knowledge graphs: lessons and challenges. *Commun. ACM* 62, 8 (2019), 36–43. doi:10.1145/3331166

[24] S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang, and X. Wu. 2024. Unifying Large Language Models and Knowledge Graphs: A Roadmap. *IEEE Trans. Knowl. Data Eng.* 36, 7 (2024), 3580–3599. doi:10.1109/TKDE.2024.3352100

[25] H. Paulheim. 2017. Knowledge Graph Refinement: A Survey of Approaches and Evaluation Methods. *Semantic Web* 8, 3 (2017), 489–508. doi:10.3233/SW-160218

[26] O. Pelgrin, R. Taelman, L. Galárraga, and K. Hose. 2025. Expressive Querying and Scalable Management of Large RDF Archives. *Semantic Web Journal* (2025). https://www.semantic-web-journal.net/system/files/swj3940.pdf

[27] A. Polleres, R. Pernisch, A. Bonifati, D. Dell'Aglio, D. Dobriy, S. Dumbrava, L. Etcheverry, N. Ferranti, K. Hose, E. Jiménez-Ruiz, M. Lissandrini, A. Scherp, R. Tommasini, and J Wachs. 2023. How Does Knowledge Evolve in Open Knowledge Graphs? *TGDK* 1, 1 (2023), 11:1–11:59. doi:10.4230/TGDK.1.1.11

[28] A. Saxena, S. Chakrabarti, and P. Talukdar. 2021. Question Answering Over Temporal Knowledge Graphs. In *Proc. of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021.* ACL, 6663–6676. doi:10.18653/V1/2021.ACL-LONG.520

[29] F. M. Suchanek, M. Alam, T. Bonald, L. Chen, P. H. Paris, and J. Soria. 2024. YAGO 4.5: A Large and Clean Knowledge Base with a Rich Taxonomy. In *Proc. of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14-18, 2024.* ACM, 131–140. doi:10.1145/3626772.3657876

[30] T. P. Tanon and F. M. Suchanek. 2019. Querying the Edit History of Wikidata. In *The Semantic Web: ESWC 2019 Satellite Events - ESWC 2019 Satellite Events, Portorož, Slovenia, June 2-6, 2019, Revised Selected Papers (LNCS, Vol. 11762).* Springer, 161–166. doi:10.1007/978-3-030-32327-1_32

[31] D. Vrandecic and M. Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Commun. ACM* 57, 10 (2014), 78–85. doi:10.1145/2629489

[32] G. Weikum, X. L. Dong, S. Razniewski, and F. M. Suchanek. 2021. Machine Knowledge: Creation and Curation of Comprehensive Knowledge Bases. *Found. Trends Databases* 10, 2-4 (2021), 108–490. doi:10.1561/1900000064

[33] B. Xue and L. Zou. 2023. Knowledge Graph Quality Management: A Comprehensive Survey. *IEEE Trans. Knowl. Data Eng.* 35, 5 (2023), 4969–4988. doi:10.1109/TKDE.2022.3150080

[34] X. Yin, J. Han, and P. S. Yu. 2008. Truth Discovery with Multiple Conflicting Information Providers on the Web. *IEEE Trans. Knowl. Data Eng.* 20, 6 (2008), 796–808. doi:10.1109/TKDE.2007.190745

[35] A. Zaveri, A. Rula, A. Maurino, R. Pietrobon, J. Lehmann, and S. Auer. 2016. Quality assessment for Linked Data: A Survey. *Semantic Web* 7, 1 (2016), 63–93. doi:10.3233/SW-150175

Preprint (author version)