Editorial of Special Issue on Knowledge Graph Construction

Journal Title
XX(X):1–9

DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

David Chaves-Fraga[1], Christophe Debruyne[2], Anastasia Dimou[3] and Maria-Esther Vidal[4]

Preface

Knowledge graphs (KGs)[1] have become essential for realizing the Semantic Web vision, enabling the development of intelligent, data-driven systems and lately supporting advancements in Large Language Models (LLMs). Yet, its construction remains a complex and evolving challenge that continues to attract attention from both academia and industry[2]. This special issue brings together recent research contributions that advance the state of the art in this field, offering theoretical insights, methodological innovations, and practical solutions that collectively push the boundaries of knowledge graphs' construction and pave the way for future innovations.

Over the past decade, significant progress has been made in developing methods and systems for transforming semi-structured data, such as tabular data (e.g., relational databases, or CSV files), or hierarchical data (e.g., XML, or JSON) into RDF-based knowledge graphs[3]. Declarative frameworks, including mapping languages[4,5] and rule-based approaches, have laid the foundations for transparent and reproducible knowledge graphs' construction[6,7]. However, important challenges remain regarding the formalization of mapping languages, the automation of mapping rules definition, the scalability and performance of knowledge graph construction systems, as well as their systematic evaluation. Addressing these issues requires both conceptual advances and efficient implementations that can cope with the increasing volume and heterogeneity of data.

[1]CiTIUS, Universidade de Santiago de Compostela, Spain
[2]Montefiore Institute, University of Liège, Belgium
[3]KU Leuven – Flanders Make@KULeuven – Leuven.AI, Belgium
[4]TIB Leibniz Information Centre for Science and Technology, Germany

Corresponding author:
David Chaves-Fraga, CiTIUS, Universidade de Santiago de Compostela, Santiago de Compostela, A Coruña 15705, Spain.
Email: david.chaves@usc.es

Prepared using sagej.cls *[Version: 2017/01/17 v1.20]*

The human factor introduces its own complexities as well, including designing intuitive and accessible user interfaces, minimizing cognitive load, and balancing automation with manual oversight. Equally demanding is the process of conducting user-centered evaluations which poses challenges in defining meaningful metrics, capturing diverse user needs, and ensuring that systems are both usable and effective in real-world scenarios. Addressing these issues requires incorporating human-in-the-loop approaches and enabling iterative feedback to ensure that the resulting systems meet real-world usability.

In parallel, the field is undergoing a transformation driven by machine learning (ML) and, more recently, large language models (LLMs)[8,9]. These approaches are opening new avenues for the automation and enrichment of knowledge graphs[10], from schema generation and entity extraction to data validation and quality assessment. While such models bring remarkable potential, they also raise new questions concerning explainability, reproducibility, and human oversight, calling for hybrid approaches that integrate declarative methods with machine learning techniques in a human-in-the-loop setting.

The contributions included in this special issue reflect these diverse but interconnected perspectives. They explore the interplay between declarative and procedural paradigms, the optimization of mapping systems, and the integration of machine learning and LLMs in the knowledge engineering lifecycle. Together, they offer a comprehensive and up-to-date picture of the research landscape, illustrating how established methodologies continue to evolve while new paradigms emerge.

This collection provides both an overview of the current state of knowledge graph construction and inspiration for further research in this vibrant and rapidly advancing area of the Semantic Web.

Contributions

This special issue received 14 articles, among which 7 were accepted for publication. Among the accepted articles, 1 article is related to mapping rules' formalization[11], 2 to optimizations of knowledge graph construction systems[12,13], 3 to large language models[14–16], 2 to the event-centric knowledge graph construction[13,16] and 1 to users' explainability of the knowledge graph construction process[17]. In detail:

- *Viktor Moskvoretskii, Irina Nikishina, Ekaterina Neminova, Alina Lobanova, Alexander Panchenko and Chris Biemann. Large Language Models for Creation, Enrichment and Evaluation of Taxonomic Graphs*[14]. This paper investigates the use of large language models for automating taxonomy-related tasks, such as construction, enrichment, and evaluation of taxonomic graphs. The authors introduce TaxoLLaMA, a unified model fine-tuned on datasets derived from WordNet 3.0, capable of addressing a broad range of lexical semantic tasks. Experimental results show that the model achieves state-of-the-art performance across multiple benchmarks, demonstrating the potential of LLMs for learning and refining taxonomic relations with high accuracy.
- *Bohui Zhang, Albert Meroño-Peñuela, and Elena Simperl. Towards Explainable Automated Knowledge Engineering with Human-in-the-loop*[17]. This paper addresses the growing opacity of modern knowledge graph construction pipelines, which increasingly depend on black-box machine learning models and heterogeneous data sources. The authors propose a framework for explainable, human-in-the-loop knowledge engineering, combining a systematic literature review with interviews from domain experts. Their study identifies key use cases and gaps in explainability, outlines user requirements, and proposes design blueprints to improve transparency, accountability, and fairness in automated KG construction processes, aligning them with emerging AI governance and regulatory frameworks.

- *Minh-Hoang Dang, Thi Hoang Thi Pham, Pascal Molli, Hala Skaf-Molli and Alban Gaignard. LLM4Schema.org: Generating Schema.org Markups with Large Language Models*[15]. This paper presents LLM4Schema.org, a novel framework for evaluating the ability of large language models to generate valid and semantically accurate Schema.org markup for web content. Unlike traditional benchmarking methods, LLM4Schema.org compares LLM-generated annotations directly with human-produced markup, without requiring predefined ground truth. The study reveals that a significant portion of LLM outputs remain invalid or non-compliant, highlighting current limitations in ontology adherence. However, the authors show that LLM-based validation agents can effectively filter and improve the quality of generated markup, with GPT-4 ultimately surpassing human performance after correction, demonstrating both the promise and the challenges of using LLMs for large-scale semantic annotation.
- *Inès Blin, Ilaria Tiddi, Remi van Trijp and Annette ten Teije. ChronoGrapher: Event-centric Knowledge Graph Construction via Informed Graph Traversal*[16]. This paper introduces ChronoGrapher, a system designed to automatically construct event-centric knowledge graphs from large, generic graphs such as DBpedia or Wikidata. The approach combines a semantically guided, best-first traversal to extract event-relevant subgraphs with a hybrid enrichment strategy that integrates structured and text-based information. ChronoGrapher demonstrates strong adaptability across datasets and outperforms existing methods in building coherent, temporally grounded event graphs. A user study further shows that incorporating event-centric triples significantly improves the factual grounding and relevance of responses in event-related question answering tasks.
- *Herminio García-González. Optimising the ShExML engine through code profiling: from turtle's pace to state-of-the-art performance*[12]. This paper focuses on improving the performance of the ShExML engine, a user-friendly language for knowledge graph construction. Through detailed code profiling and optimization, the author identifies and resolves key performance bottlenecks, followed by a rigorous statistical evaluation of the enhancements. The optimized engine achieves execution times comparable to other state-of-the-art mapping systems, significantly improving its scalability and responsiveness. As a result, ShExML now provides a more efficient and reliable solution for users constructing knowledge graphs from heterogeneous data sources.
- *Sitt Min Oo, Ben De Meester, Ruben Taelman and Pieter Colpaert. Algebraic Mapping Operators for Knowledge Graph Generation*[11]. This paper introduces a formal algebraic framework for defining the operational semantics of knowledge graph generation processes. Building upon the SPARQL algebra, the proposed algebraic mapping operators provide a unified foundation for representing and optimizing mappings across different languages, such as RML and ShExML. The authors demonstrate that this approach enables language-independent optimization and fosters consistency between mapping and query engines. Experimental results show stable, low-memory performance, with the prototype ranking highly in the Knowledge Graph Construction Workshop's performance challenge. This work establishes a solid theoretical basis for analyzing the complexity, expressiveness, and optimization of knowledge graph mapping systems.
- *Dylan Van Assche, Julian Rojas, Ben De Meester and Pieter Colpaert. Incremental Knowledge Graph Construction from Heterogeneous Data Sources*[13]. This paper presents IncRML, an incremental approach to knowledge graph generation that efficiently handles evolving and dynamic data sources. Built on top of RML and FnO, the method detects and materializes

data changes—creates, updates, and deletes—while minimizing redundant processing. Detected changes can be published as Linked Data Event Streams (LDES) using the W3C Activity Streams 2.0 vocabulary, enabling semantic communication of updates over the Web. Experimental results across multiple real-world datasets demonstrate substantial reductions in storage, CPU time, and memory usage, achieving up to 315× lower storage costs and 4.4× faster construction. The approach significantly lowers the cost of maintaining up-to-date knowledge graphs and promotes scalable, Web-native publication of evolving data.

Analysis and beyond

After analyzing the accepted articles, their main contributions focus on two areas: the development of standards and formalization techniques for declarative methods for knowledge graph construction, and the automation or provision of user support throughout the knowledge graph construction process.

Standards and Standardization

A recurring theme in knowledge graph construction is the role of standards and their adoption by the community. In this special issue, we received contributions related to the RML[4,5] and ShexML[18] mapping languages. RML extends R2RML[*], a W3C Recommendation for transforming data contained in relational databases into RDF. RML has been instrumental in supporting mappings beyond relational databases. The Knowledge Graph Construction Community Group[†] was established in 2019 and has been actively working toward establishing a charter for a candidate submission to the W3C since 2023. Similarly, ShExML builds upon ShEx[‡], which, although not a W3C Recommendation, is widely acknowledged as a specification for validation by the Semantic Web community. This close alignment with recognized and widely adopted specifications underscores how ShExML, like RML, is rooted within standards, whether identified by a standardization body or community.

We note that this special issue did not receive contributions on Façade-X[19,20]. Yet, it is important to recognize that a W3C Community Group[§] was established in September 2025 for this initiative as well. Façade-X enables the use of SPARQL to interact with various sources as RDF graphs. RML and ShExML differ from Façade-X in that the former focus on declaratively specifying mappings from source data to RDF according to a target ontology. In contrast, the latter focuses on using SPARQL to access data contained in heterogeneous data sources as RDF via direct mappings. These direct mappings are a natural evolution of a direction mapping approach as proposed by the *Direct Mapping of Relational Data to RDF*[¶] and generalized to different data sources by Façade-X. With Façade-X, one can thus use SPARQL CONSTRUCT to generate RDF graphs using different vocabularies and create knowledge graph construction pipelines.

Taken together, these developments suggest that emerging standards are evolving to address different use cases and niches in knowledge graph construction. However, the different approaches highlight

[*] https://www.w3.org/TR/r2rml/
[†] https://www.w3.org/community/kg-construct/
[‡] https://shex.io/
[§] https://www.w3.org/groups/cg/fx/
[¶] https://www.w3.org/TR/rdb-direct-mapping/

both the richness of the field and the importance of continued dialogue between communities to ensure maximum interoperability and sustainable adoption.

Mapping Languages' Formalization

Another observation is the increased effort within the community to formalize the mapping languages for the knowledge graph construction. While R2RML introduced a reference algorithm, it did not provide a formal foundation. This gap has been addressed by, for example, [21] and [22]. The first offered a formalization of R2RML for the Ontop system; the second proposed a Datalog-based formalization of R2RML. In this special issue, Min Oo et al. presented initial steps toward formalizing RML. Although this work was well received, we note for the readers' benefit that the lead author of the article has recently published a similar work in collaboration with a computer scientist [23], which won the Best Paper award at the 22nd European Semantic Web Conference (ESWC) in 2025.

We expect that the formalization of mapping languages will play an important role in shaping and driving the standardization process of knowledge graph construction. Historically, knowledge graph construction has been led primarily by engineers focused on system development and practical applications. In contrast, formalization is often the domain of (theoretical) computer scientists. As KGC languages grow more expressive and begin to intersect with fields such as programming languages, logic, and database theory, closer collaboration between these communities will be essential. It will otherwise be challenging to establish solid, extensible, and verifiable foundations.

Large Language Models

A notable trend is the integration of large language models (LLMs) to streamline and enhance various stages of knowledge graph construction, ranging from initial design to iterative refinement and enrichment. This relationship between Large Language Models (LLMs) and knowledge graphs is bidirectional. On one hand, LLMs can support and accelerate the construction of knowledge graphs; on the other hand, the continuous enhancement of knowledge graphs with newly constructed knowledge strengthens the performance of LLMs, improving their factual grounding, reducing hallucinations, and enabling more accurate reasoning. LLMs have been considered so far for all aspects of knowledge graph construction; ranging from ontology development and taxonomy construction and enrichment, as with [14], to table interpretation and knowledge graph enrichment. However, as automated approaches increasingly rely on LLMs, challenges arise regarding the validity and interpretability of the results, as [15] showed. Ambiguous or incorrect results highlights the need for mechanisms for their explainability, as with [17], validation, as with [15], and improvement are required.

Moreover, as knowledge graph construction is a complex process that requires good understanding of the data to semantically annotate entities and their relations, we observe that knowledge graph construction with declarative mapping languages remains primarily manual. Automated systems for the construction of knowledge graphs rely more and more on LLMs, but either only indicate the semantic annotations without constructing a knowledge graph, or directly produce a knowledge graph without producing any declarative mapping rules. As with declarative knowledge graph construction communities that have their own community groups and venues, e.g., the knowledge graph construction

workshop[‖], the automated knowledge graph construction topic has its own communities, e.g., around the SemTab challenge[**] or the Table Representation Learning Workshop[††]. Consequently, we observe that the declarative and automated approaches evolve in different directions and little interaction occurs between them.

Human-in-the-Loop

Although human interactions for knowledge graph construction was one of the highlighted topics of the special issue, only one article addressed this aspect. Existing approaches for human interactions tend to fall into two categories: declarative methods, which emphasize editor tools and structured methodologies such as the pay-as-you-go paradigm[24] to assist users, and automated solutions that often rely on human feedback for refinement. However, there has been limited investment in comprehensive human-in-the-loop methodologies that actively integrate user input throughout the entire construction process. Recognizing this gap, a dedicated workshop on "Users and Knowledge Graphs"[‡‡] was recently organized to foster dialogue, share best practices, and promote research that prioritizes usability, collaboration, and human-centered design in knowledge graph development.

Challenges and Future Work

Knowledge graph construction (KGC) continues to evolve rapidly, driven by new paradigms, technologies, and the increasing need for scalable, explainable, and interoperable data integration methods. Despite the significant advances reflected in this special issue, several open challenges remain and point toward promising directions for future research.

A first challenge lies in bridging declarative and learning-based approaches. Declarative mapping languages provide transparency, reproducibility, and formal rigor, whereas machine learning and LLMs offer adaptability and automation. The integration of both paradigms, through neuro-symbolic approaches with human-in-the-loop pipelines, remains an open area that requires methodological frameworks and evaluation metrics capable of balancing interpretability and performance.

A second line of research concerns evaluation and benchmarking. Although several benchmarks such as KROWN[25] and GTFS-Madrid-Bench[26] have played an important role in assessing the performance of KGC engines, they require updates to align with current languages, specifications and parameters. Moreover, the rise of LLM-based and incremental construction approaches calls for new evaluation scenarios that go beyond traditional materialization pipelines. Recent initiatives such as the BLINKG benchmark[27] represent a significant step toward reproducible, fine-grained, and comparable evaluation of automatic KGC methods. However, broader community coordination is still needed to establish unified datasets, metrics, and evaluation protocols that ensure fairness and cumulative scientific progress.

Another emerging challenge involves maintaining and evolving KGs in dynamic and distributed environments. The growth of data spaces and the proliferation of real-time data streams call for incremental and event-driven graph construction methods, enabling continuous synchronization,

[‖] http://w3id.org/kg-construct/workshop
[**] https://www.cs.ox.ac.uk/isg/challenges/sem-tab/
[††] https://sites.google.com/view/eurips25-ai-td/home
[‡‡] https://ukgworkshop.github.io

provenance tracking, and version management while minimizing computational costs[28]. Finally, from a foundational perspective, formalization and standardization remain essential. Ongoing efforts toward algebraic semantics for mapping languages and formal models of execution need to be consolidated[23] and connected with W3C standardization initiatives. Ensuring semantic interoperability across languages, engines, and specifications will be key to achieving maturity and long-term sustainability in the field.

In summary, the future of knowledge graph construction lies in combining solid theoretical foundations with adaptive, explainable, and interoperable systems. The convergence of declarative design, formal semantics, and machine learning holds the potential to make KGC not only more efficient and scalable but also more trustworthy and impactful across scientific and industrial domains.

References

1. Hogan A, Blomqvist E, Cochez M et al. Knowledge graphs. *ACM Computing Surveys (Csur)* 2021; 54(4): 1–37.
2. Chaves-Fraga D, Corcho O, Dimou A et al. Are knowledge graphs ready for the real world? challenges and perspective. *Dagstuhl Reports* 2024; 14(2): 1–70.
3. Van Assche D, Delva T, Haesendonck G et al. Declarative rdf graph generation from heterogeneous (semi-) structured data: A systematic literature review. *Journal of Web Semantics* 2023; 75: 100753.
4. Iglesias-Molina A, Van Assche D, Arenas-Guerrero J et al. The RML ontology: A community-driven modular redesign after a decade of experience in mapping heterogeneous data to RDF. In Payne TR, Presutti V, Qi G et al. (eds.) *The Semantic Web - ISWC 2023 - 22nd International Semantic Web Conference, Athens, Greece, November 6-10, 2023, Proceedings, Part II*, *Lecture Notes in Computer Science*, volume 14266. Springer, pp. 152–175. DOI:10.1007/978-3-031-47243-5_9. URL https://doi.org/10.1007/978-3-031-47243-5_9.
5. Dimou A, Vander Sande M, Colpaert P et al. RML: A generic language for integrated RDF mappings of heterogeneous data. In Bizer C, Heath T, Auer S et al. (eds.) *Proceedings of the Workshop on Linked Data on the Web co-located with the 23rd International World Wide Web Conference (WWW 2014), Seoul, Korea, April 8, 2014*, *CEUR Workshop Proceedings*, volume 1184. CEUR-WS.org. URL https://ceur-ws.org/Vol-1184/ldow2014_paper_01.pdf.
6. Iglesias E, Jozashoori S, Chaves-Fraga D et al. Sdm-rdfizer: An rml interpreter for the efficient creation of rdf knowledge graphs. In *Proceedings of the 29th ACM international conference on Information & Knowledge Management*. pp. 3039–3046.
7. Arenas-Guerrero J, Chaves-Fraga D, Toledo J et al. Morph-kgc: Scalable knowledge graph materialization with mapping partitions. *Semantic Web* 2024; 15(1): 1–20.
8. Schmidt WJ, Grangel-González I, Huschle T et al. Llm-supported mapping generation for semantic manufacturing treasure hunting. In *European Semantic Web Conference*. Springer, pp. 84–101.
9. Freund M, Dorsch R, Schmid S et al. Mapping by example: Towards an rml mapping reverse engineering pipeline. In *Sixth International Workshop on Knowledge Graph Construction@ ESWC2025*.
10. Dimou A and Chaves-Fraga D. Declarative description of knowledge graphs construction automation: Status & challenges. In *Proceedings of the 3rd International Workshop on Knowledge Graph Construction (KGCW 2022) co-located with 19th Extended Semantic Web Conference (ESWC 2022)*, volume 3141.
11. Oo SM, De Meester B, Taelman R et al. Algebraic mapping operators for knowledge graph generation. *Semantic Web* 2025; 16(5): 22104968251361350. DOI:10.1177/22104968251361350. URL https://doi.org/10.1177/22104968251361350.

12. García-González H. Optimising the shexml engine through code profiling: From turtle's pace to state-of-the-art performance. *Semantic Web* 2025; 16(2): SW–243736. DOI:10.3233/SW-243736. URL https://journals.sagepub.com/doi/abs/10.3233/SW-243736.
13. Van Assche D, Rojas J, De Meester B et al. Incremental knowledge graph construction from heterogeneous data sources. *Semantic Web* 2025; .
14. Moskvoretskii V, Neminova E, Lobanova A et al. Large language models for creation, enrichment and evaluation of taxonomic graphs. *Semantic Web* 2025; .
15. Dang MH, Pham THT, Molli P et al. Llm4schema. org: Generating schema. org markups with large language models. *Semantic Web* 2025; .
16. Blin I, Tiddi I, van Trijp R et al. Chronographer: Event-centric knowledge graph construction via informed graph traversal. *Semantic Web* 2025; 16(5): 22104968251377247. DOI:10.1177/22104968251377247. URL https://doi.org/10.1177/22104968251377247.
17. Zhang B, Meroño-Peñuela A and Simperl E. Towards explainable automated knowledge engineering with human-in-the-loop. *Semantic Web* 2025; 16(5): 22104968251382171. DOI:10.1177/22104968251382171. URL https://doi.org/10.1177/22104968251382171.
18. García-González H, Boneva I, Staworko S et al. Shexml: improving the usability of heterogeneous data mapping languages for first-time users. *PeerJ Comput Sci* 2020; 6: e318. DOI:10.7717/PEERJ-CS.318. URL https://doi.org/10.7717/peerj-cs.318.
19. Daga E, Asprino L, Mulholland P et al. Facade-x: An opinionated approach to sparql anything. In Alam M, Groth P, de Boer V et al. (eds.) *Further with Knowledge Graphs - Proceedings of the 17th International Conference on Semantic Systems, SEMANTiCS 2021, Amsterdam, The Netherlands, September 6-9, 2021*, *Studies on the Semantic Web*, volume 53. IOS Press, pp. 58–73. DOI:10.3233/SSW210035. URL https://doi.org/10.3233/SSW210035.
20. Asprino L, Daga E, Dowdy J et al. Materialisation approaches for façade-based data access with sparql. *Semantic Web* 2024; URL https://www.semantic-web-journal.net/content/materialisation-approaches-fa%C3%A7ade-based-data-access-sparql-0. Accepted for publication.
21. Calvanese D, Cogrel B, Komla-Ebri S et al. Ontop: Answering SPARQL queries over relational databases. *Semantic Web* 2017; 8(3): 471–487. DOI:10.3233/SW-160217. URL https://doi.org/10.3233/SW-160217.
22. Elhalawati A, Jan Van den Bussche and Dimou A. A declarative formalization of r2rml using datalog and its efficient execution. In Margara A, Kliegr T, Savkovic O et al. (eds.) *Companion Proceedings of the 9th International Joint Conference on Rules and Reasoning (RuleML+RR 2025) also co-located with 21th Reasoning Web Summer School (RW 2025) and 17th DecisionCAMP 2025 as part of Declarative AI 2025, Istanbul, Türkiye, September 22-24, 2025.*, *CEUR Workshop Proceedings*, volume 4083. CEUR-WS.org. URL https://ceur-ws.org/Vol-4083/paper63.pdf.
23. Oo SM and Hartig O. An algebraic foundation for knowledge graph construction. In Curry E, Acosta M, Poveda-Villalón M et al. (eds.) *The Semantic Web - 22nd European Semantic Web Conference, ESWC 2025, Portoroz, Slovenia, June 1-5, 2025, Proceedings, Part I*, *Lecture Notes in Computer Science*, volume 15718. Springer, pp. 3–22. DOI:10.1007/978-3-031-94575-5_1. URL https://doi.org/10.1007/978-3-031-94575-5_1.
24. Sequeda JF and Miranker DP. A pay-as-you-go methodology for ontology-based data access. *IEEE Internet Computing* 2017; 21(2): 92–96.

25. Van Assche D, Chaves-Fraga D and Dimou A. Krown: A benchmark for rdf graph materialisation. In *International Semantic Web Conference*. Springer, pp. 20–39.
26. Chaves-Fraga D, Priyatna F, Cimmino A et al. Gtfs-madrid-bench: A benchmark for virtual knowledge graph access in the transport domain. *Journal of Web Semantics* 2020; 65: 100596.
27. Castedo C, Iglesias E, Lama M et al. Blinkg: A benchmark for llm-integrated knowledge graph generation. *Transactions on Graph Data and Knowledge* 2026; Under Review.
28. Geisler S, Cappiello C, Celino I et al. From genesis to maturity: managing knowledge graph ecosystems through life cycles. *Proceedings of the VLDB Endowment* 2025; 18(5): 1390–1397.