Semantic Web 0 (2024) 1–0
IOS Press

Special Issue on Wikidata Construction, Evaluation and Applications

Editorial

Lucie Kaffee [a,*],
[a] *HuggingFace, Germany*
E-mail: lucie.kaffee@gmail.com
Simon Razniewski [b] and
[b] *ScaDS.AI & TU Dresden, Germany*
E-mail: simon.razniewski@tu-dresden.de
Pavlos Vougiouklis [c]
[c] *Huawei Technologies, UK*
E-mail: pavlos.vougiouklis@huawei.com

Abstract. Wikidata [3], the open knowledge graph maintained by the Wikimedia Foundation, continues to expand its role as a central hub of structured data for Wikipedia and its sister projects, as well as an increasingly vital resource for academic research and industrial applications. Its collaborative nature, blending human and machine contributions, creates unique opportunities and challenges at the intersection of knowledge representation, data management, and socio-technical systems. The growing academic interest in Wikidata has been exemplified by initiatives like the annual Wikidata Workshop at the International Semantic Web Conference (ISWC) [1, 2]. These efforts have fostered a vibrant research community investigating both the technical aspects of Wikidata and its socio-technical ecosystem. Building on this momentum, this special issue of the Semantic Web Journal brings together cutting-edge research to explore the opportunities and challenges posed by this collaborative knowledge graph.

Keywords: keywords

1. Preface

Wikidata [3], the open knowledge graph maintained by the Wikimedia Foundation, continues to expand its role as a central hub of structured data for Wikipedia and its sister projects, as well as an increasingly vital resource for academic research and industrial applications. Its collaborative nature, blending human and machine contributions, creates unique opportunities and challenges at the intersection of knowledge representation, data management, and socio-technical systems.

The growing academic interest in Wikidata has been exemplified by initiatives like the annual Wikidata Workshop at the International Semantic Web Conference (ISWC) [1, 2]. These efforts have fostered a vibrant research community investigating both the technical aspects of Wikidata and its socio-technical ecosystem. Building on this momentum, this special issue of the Semantic Web Journal brings together cutting-edge research to explore the opportunities and challenges posed by this collaborative knowledge graph.

*Corresponding author. E-mail: lucie.kaffee@gmail.com.

The articles in this issue address a range of critical themes. Some contributions focus on data quality and validation, offering frameworks and tools for assessing referencing practices and formalizing property constraints. These studies emphasize the importance of reliable data for both the immediate Wikidata community and its broader ecosystem of users. Other articles delve into the structure and representation of knowledge within Wikidata, uncovering patterns and inconsistencies in its ontologies and taxonomies that reflect the complexities of a community-driven knowledge graph.

Beyond these foundational topics, several contributions demonstrate the innovative applications made possible by Wikidata. For instance, researchers explore how it can support cultural heritage storytelling, enhance the representation of scientific knowledge, and enable natural language generation through its lexemes and items. At the same time, the issue highlights the importance of trust and provenance, showcasing methodologies for verifying data against textual sources and examining the factors that influence user confidence in Wikidata's content.

The socio-technical dimensions of Wikidata also receive significant attention, with research examining how its collaborative dynamics influence knowledge creation and curation. Studies in this area shed light on the challenges of managing diverse contributions, as well as the potential for automated tools to complement human efforts. Finally, the issue features work that advances practical solutions, such as subsetting tools to manage Wikidata's vast scale and frameworks for evaluating its performance in question-answering tasks.

As editors, we are inspired by the depth and breadth of the work presented in this special issue. We hope it serves as a valuable resource for researchers, practitioners, and members of the Wikidata community, fostering further innovation and collaboration. With continued investment from academia, industry, and the community, Wikidata will undoubtedly remain a critical resource for advancing open knowledge on the web.

2. Topics of Interest

- Data quality and vandalism detection in Wikidata
- Referencing in Wikidata
- Anomaly, bias, or novelty detection in Wikidata
- Algorithms for aligning Wikidata with other structured or semi-structured resources
- Representation, Semantic Annotation, Enhancement and Enrichments using Wikidata
- Semantic Parsing and Question Answering using Wikidata
- The Semantic Web and Wikidata
- Community interaction in Wikidata
- Multilingual data in Wikidata and its reuse
- Data quality in Wikidata: Approaches for problem detection, evaluation and improvement
- Tools, bots, and datasets for improving or evaluating Wikidata
- Participation, diversity, and inclusivity aspects in the Wikidata ecosystem
- Human-bot interaction
- Managing knowledge evolution in Wikidata
- Abstract Wikipedia
- Wikidata in NLP applications

3. Content

This issue consists of the following 13 papers:

1. RQSS: Referencing Quality Scoring System for Wikidata
2. On assessing weaker logical status claims in Wikidata Cultural Heritage records
3. InteractOA: Showcasing the representation of knowledge from scientific literature in Wikidata
4. Empirical ontology design patterns and shapes from Wikidata
5. Formalizing and Validating Wikidata's Property Constraints using SHACL and SPARQL

6. Can you trust Wikidata?
7. Using Wikidata Lexemes and Items to Generate Text from Abstract Representations
8. Evidence of Large-Scale Conceptual Disarray in Multi-Level Taxonomies in Wikidata
9. Dura-Europos Stories: Developing Interactive Storytelling Applications Using Knowledge Graphs for Cultural Heritage Exploration
10. Wikidata subsetting: approaches, tools, and evaluation
11. QALD-10 — The 10th Challenge on Question Answering over Linked Data
12. ProVe: A Pipeline for Automated Provenance Verification of Knowledge Graphs Against Textual Sources
13. Psychiq and Wwwyzzerdd: Wikidata completion using Wikipedia

In the following, we introduce their topics:

In *"RQSS: Referencing Quality Scoring System for Wikidata"* by *Seyed Amir Hosseini Beghaeiraveri, Alasdair J G Gray, and Fiona McNeill*, the authors present a comprehensive framework, the Referencing Quality Scoring System (RQSS), to evaluate the quality of references in Wikidata based on Linked Data quality dimensions. The study demonstrates RQSS's effectiveness in analyzing referencing quality across various topical and random subsets, revealing strengths in dimensions like accuracy and availability but identifying gaps in completeness and verifiability, with an overall referencing quality score of 0.58.

In *"On assessing weaker logical status claims in Wikidata Cultural Heritage records"* by *Alessio Di Pasquale, Valentina Pasqual, Francesca Tomasi, and Fabio Vitali*, the authors investigate how Wikidata represents uncertain or evolving information using approaches like non-asserted statements, ranked statements, and specific qualifiers. The study reveals that while weaker logical status information is limited in prevalence, its representation is often ambiguous, prompting recommendations for simplifying and standardizing these practices to enhance clarity and usability.

In *"InteractOA: Showcasing the representation of knowledge from scientific literature in Wikidata"* by *Muhammad Elhossary and Konrad Foerstner*, the authors introduce InteractOA, a tool for integrating knowledge about small regulatory RNAs (sRNAs) into Wikidata, linking it directly to its origins in open-access scholarly articles. This system demonstrates how life science findings can be stored in a machine-readable, queryable format within a rich knowledge graph, encouraging researchers to contribute, edit, and reuse structured scientific data.

In *"Empirical Ontology Design Patterns and Shapes from Wikidata"* by *Valentina Anita Carriero, Paul Groth, and Valentina Presutti*, the authors propose a method for identifying empirical ontology design patterns (EODPs) to uncover and formalize the implicit ontology of Wikidata. By analyzing domain-specific subsets of Wikidata in music and art-related fields, they demonstrate how these patterns can guide ontology usage, improve its structure, and provide insights into domain-specific knowledge graph content.

In *"Formalizing and Validating Wikidata's Property Constraints using SHACL and SPARQL"* by *Nicolas Ferranti, Jairo Francisco de Souza, Shqiponja Ahmetaj, and Axel Polleres*, the authors explore the use of SHACL and SPARQL to formalize and validate Wikidata's property constraints, enhancing data integrity. They demonstrate that while SHACL-Core lacks the expressivity to cover all 32 Wikidata constraint types, SPARQL queries can effectively identify violations, offering insights into improving constraint management and contributing to benchmarks for large-scale knowledge graph validation.

In *"Can You Trust Wikidata?"* by *Veronica Santos, Daniel Schwabe, and Sérgio Lifschitz*, the authors investigate Wikidata's ability to support trust decisions when using its data, particularly in the context of its crowdsourced nature and diverse information sources. Through Knowledge Graph profiling, they analyze how mechanisms like claims, schemas, and representations of multiple viewpoints and controversies contribute to assessing the veracity and reliability of Wikidata's content.

In *"Using Wikidata Lexemes and Items to Generate Text from Abstract Representations"* by *Mahir Morshed*, the author introduces Ninai/Udiron, a natural language generation system that transforms abstract factual representations into human-readable text using Wikidata lexemes and items. Designed for efficiency, extensibility, and analyzability, the system generates syntax trees and sentences, aiming to support the Abstract Wikipedia project and be integrated into the Wikifunctions platform.

In *"Evidence of Large-Scale Conceptual Disarray in Multi-Level Taxonomies in Wikidata"* by *Atílio A. Dadalto, João Paulo A. Almeida, Claudenir M. Fonseca, and Giancarlo Guizzardi*, the authors examine widespread issues

in distinguishing between types and individuals in Wikidata's taxonomies, resulting in frequent modeling errors. They propose methodological and computational tools to address these issues, offering a conceptual analysis and a supportive tool to help mitigate errors in representing instantiation and specialization relationships.

In *"Dura-Europos Stories: Developing Interactive Storytelling Applications Using Knowledge Graphs for Cultural Heritage Exploration"* by *Katherine Thornton, Kenneth Seals-Nutt, and Anne Chen*, the authors present a multimedia application that uses Wikidata to display artifacts and locations from the Dura-Europos archaeological excavation. The application combines visual exploration with metadata, leveraging the Semantic Web and knowledge graphs to create an interactive storytelling experience that connects users to related cultural heritage data.

In *"Wikidata Subsetting: Approaches, Tools, and Evaluation"* by *Seyed Amir Hosseini Beghaeiraveri, Jose Emilio Labra-Gayo, Andra Waagmeester, Ammar Ammar, Carolina Gonzalez, Denise Slenter, Sabah Ul-Hasan, Egon L. Willighagen, Fiona McNeill, and Alasdair J G Gray*, the authors explore various tools and approaches for subsetting Wikidata to make it more manageable for research purposes. They evaluate four tools—WDSub, KGTK, WDumper, and WDF—highlighting their performance, accuracy, and flexibility, and demonstrate how subsetting enables more targeted data extraction from Wikidata, offering valuable insights that cannot be obtained via the public SPARQL endpoint.

In *"QALD-10 — The 10th Challenge on Question Answering over Linked Data"* by *Ricardo Usbeck, Xi Yan, Aleksandr Perevalov, Longquan Jiang, Julius Schulz, Angelie Kraft, Cedric Moeller, Junbo Huang, Jan Reineke, Axel-Cyrille Ngonga Ngomo, Muhammad Saleem, and Andreas Both*, the authors present a new multilingual, complex benchmarking dataset for Knowledge Graph Question Answering (KGQA) based on Wikidata, as part of the QALD challenge series. They discuss the migration process from DBpedia to Wikidata, the challenges faced in mapping languages and properties, and how their updated benchmark aims to foster advances in KGQA by providing more demanding and diverse evaluation tasks.

In *"ProVe: A Pipeline for Automated Provenance Verification of Knowledge Graphs against Textual Sources"* by *Gabriel Amaral, Odinaldo Rodrigues, and Elena Simperl*, the authors present ProVe, a pipeline designed to automatically verify the provenance of knowledge graph triples against textual sources. Evaluated on a Wikidata dataset, ProVe uses rule-based methods and machine learning to assess whether triples are supported by their documented provenance, achieving high accuracy and F1 scores, and providing a scalable solution for ensuring the trustworthiness of knowledge graph data.

In *"Psychiq and Wwwyzzerdd: Wikidata Completion Using Wikipedia"* by *Daniel Erenrich*, the author introduces two practical tools to address the incompleteness and inaccuracies in Wikidata. Wwwyzzerdd, a browser extension, facilitates the rapid import of statements from Wikipedia to Wikidata, while Psychiq is a model designed to predict instance and subclass statements from English Wikipedia articles, contributing to the acceleration of Wikidata's completion process.

References

[1] L.-A. Kaffee, S. Razniewski, G. Amaral and K.S. Alghamdi (eds), Proceedings of the 3rd Wikidata Workshop, CEUR Workshop Proceedings, 2023. https://ceur-ws.org/Vol-3262/.

[2] L.-A. Kaffee, S. Razniewski, K. Alghamdi and H. Arnaout (eds), Proceedings of the 4th Wikidata Workshop, CEUR Workshop Proceedings, 2024. https://ceur-ws.org/Vol-3640/.

[3] D. Vrandečić and M. Krötzsch, Wikidata: a free collaborative knowledgebase, *Communications of the ACM* **57**(10) (2014), 78–85.