Selected Papers of #AoIR2025:
The 26th Annual Conference of the
Association of Internet Researchers
Niterói, Brazil / 15 – 18 Oct 2025

WIKIDATA'S WORLDVIEW: INSPECTING AN AI KNOWLEDGE PIPELINE WITH SEMANTIC NETWORK ANALYSIS

Andrew Iliadis
Temple University

Mikayla Gonzalez
Temple University

Introduction

Wikidata is a collaboratively edited, open knowledge graph (i.e., a network of entities and relationships used for structuring and linking information) and serves as a primary source of machine-readable data (Vrandečić & Krötzsch, 2014). Created in 2012 to aid Wikipedia with updating information across all language versions, today's Wikidata assists products like Google's Knowledge Graph and Amazon's Alexa by offering free to-use structured data in search results. As researchers have shown (Iliadis, 2022; Ford, 2022; Iliadis & Ford, 2023; Ford & Iliadis, 2023), Wikidata's influence is not limited to data retrieval but profoundly affects the way data is surfaced, organized, and transformed into computable knowledge for consumers using search products, such as Google Search, Bing, or chatbots. This Wikidata information often appears in "info boxes" (e.g., Google knowledge panels), and several search studies have shown that people trust the information that they find there (Lurie & Mustafaraj, 2018; Masullo et al., 2024; Ray, 2020; Rothschild et al., 2019). This paper uses semantic network analysis to uncover and analyze Wikidata's underlying ontology (i.e., its data structure and terminology), which is hidden from everyday web and Wikipedia users.

Literature Review

While Wikidata has become a crucial resource for structured knowledge representation, research suggests improvements in multilingual inclusivity, data quality, and knowledge integration are necessary for its continued evolution. Studies emphasize the need for enhanced provenance tracking (Santos et al., 2024), better constraint enforcement mechanisms (Shenoy et al., 2022), and more structured approaches to discussion and collaboration (Koutsiana et al., 2024). Ford and Iliadis (2023) stress the need to critically assess Wikidata's infrastructural role in shaping digital epistemologies and its implications for knowledge equity. As Wikidata continues to influence AI applications

Suggested Citation (APA):Iliadis, A., & Gonzalez, M. (2025, October). *Wikidata's Worldview: Inspecting an AI Knowledge Pipeline with Semantic Network Analysis*. Paper presented at AoIR2025: The 26th Annual Conference of the Association of Internet Researchers.Niterói, Brazil: AoIR. Retrieved from http://spir.aoir.org.

and knowledge-based systems (Vrandečić et al., 2023), addressing these challenges

will be vital to understanding its mediating role in shaping human-machine communication and information retrieval (Guzman & Lewis, 2020).

Despite its extensive use in structuring machine-readable knowledge, there is still much to learn about how Wikidata's ontology – the structured system that defines how different pieces of information relate to each other, like a big map connecting facts – shapes machine communication and AI. Its classification system influences how information is encoded, retrieved, and interpreted, yet the underlying structure, terminology, and potential biases remain underexplored. Finding inspiration in Graham et al.'s (2015) study of Wikipedia, which rejected the so-called "view from nowhere" (Nagel, 1986) by examining highly uneven geographies of participation in the encyclopedia, we wanted to reject the notion that Wikidata is neutral or objective in its description of the world and to see instead if the database structure has something like a worldview (i.e., a perspective through which to interpret and understand life and reality). We thus pose the following three research questions in our work:

RQ1: *What is Wikidata's ontological structure?*
RQ2: *What terminology does the ontology use?*
RQ3: *What sociocultural biases are visible in the ontology?*

Methods

We followed a mixed-methods research design to explore the ontological structure, terminology, and potential sociocultural biases of Wikidata. The methodology incorporates ontology network analysis (Figueres-Esteban et al., 2016; Weng et al., 2008) with semantic network analysis (Segev, 2022) and uses computational methods to visualize and extract data. The study draws upon data stored in triples, the basic form of storing information in the Resource Description Framework (RDF). A triple consists of a subject, predicate, and object, which create relationships between different entities when combined. To extract data, we utilized the Wikidata Query Service (WDQS) to run a customized structured data retrieval through the SPARQL Protocol and RDF Query Language (SPARQL). We used the Wikidata Query Builder (WQB) to create structured queries and search for entities and relationships in Wikidata. Modifications of this query were used to pull data relating to subclasses at lower levels of the ontology representing controversial or social topics. Gephi software was used to visualize the output using the Yifan Hu (Hu, 2005) and Fruchterman-Reingold (Fruchterman & Reingold, 1991) graph layout algorithms. The network's modularity was measured using Gephi's statistical tools to identify communities (i.e., clusters of topics) within the knowledge graph, utilizing the Louvain method for community detection (Blondel et al., 2008).

Preliminary Findings

Concerning RQ1 (*ontological structure*), upon analyzing the ontological structure of Wikidata, Wikidata's knowledge base is found to be overwhelmed with a myriad of concepts and relations due to its extreme complexity. One important finding is how Wikidata integrates a hierarchical classification system that enables entities to be

connected through a structured ontology capable of efficient information retrieval. Despite this, some classifications showed poor standards due to category overlaps and vague entity allocations. Different entities were found to represent a single concept in multiple ways, creating conflicts with knowledge representation (e.g., there are two entities for Justice). The existence of such inconsistencies can make it hard to achieve uniformity in interpreting and using data across domains. Figure 1 provides the topmost semantic network visualization of Wikidata's upper-level ontology. Figure 2 (top) shows the modularity of the main eighteen community groups in the topmost upper level of the ontology and how many sub-entities are connected to them. Figure 2 (bottom) shows a semantic network visualization of those communities. The main entities representing those communities in this upper level of the Wikidata ontology are Entity, Property, Part, Class, Object, Result, Former Entity, Copy, Hypothetical Entity, Source, Foundational Model of Anatomy, Attribute Entity, Union, Non-Existent Entity, Continuant, Role, Unknown, and Collective Entity.

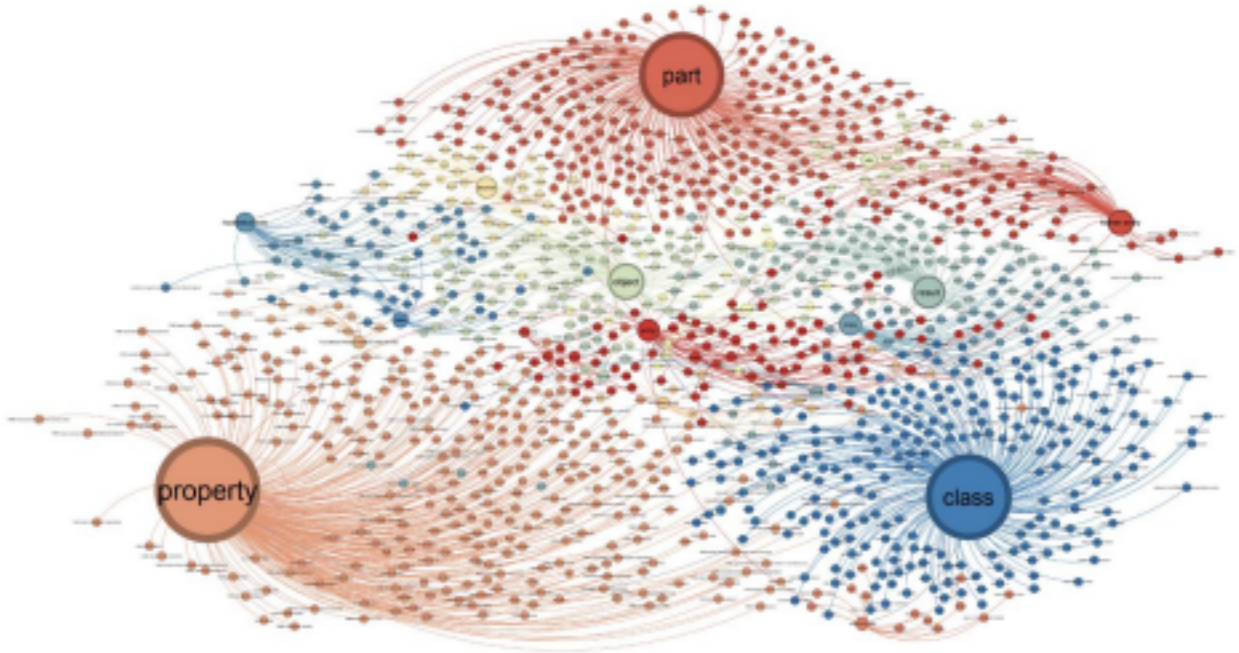

Figure 1. Semantic network visualization of Wikidata's upper-level ontology.

Concerning RQ2 (*terminology*), one of the most startling findings of the study was that Wikidata contains interestingly vague or subjective terminology located within the ontology. It was noted that terms such as Pricelessness, Greatness, and Acceptability are defined and connected to a plurality of entities in the ontology. While sitting in a well-defined database, these terms lacked definitive meanings that could be agreed upon, bringing forth concerns regarding the ontological neutrality and objectivity of Wikidata. The study also found discrepancies in how negative and controversial concepts were labeled. For instance, categorizing Bad includes anything from environmental dangers to fictional characters, which shows a very loose application of the term. These loose definitions indicate that the sociocultural biases of Wikidata editors influence and shape how Wikidata organizes negative or contested concepts.

Concerning RQ3 (*sociocultural biases*), a lack of precision can harm semantic search

engines and digital assistants that depend on well-defined data. Algorithms that fetch information from Wikidata may also reinforce the biases existing in the ontology. Take, for example, the Notable Work classification. The term is primarily associated with Western literature, which excludes underrepresented cultures and thus creates a false narrative regarding what is important in the history of culture and literature. The same goes for Employment, where specific job titles are affiliated with certain genders or races due to sociocultural preconceptions ingrained in the data.

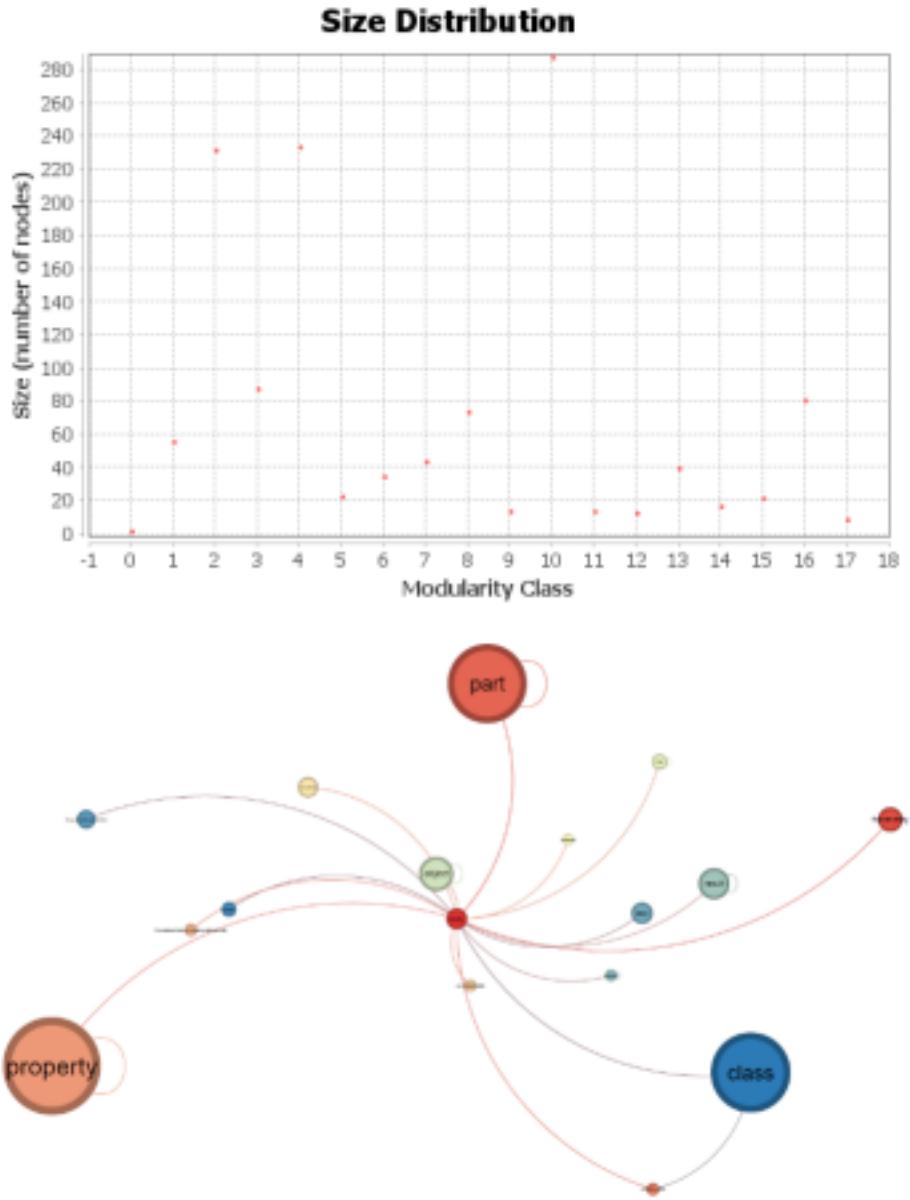

Figure 2. Modularity statistics for Wikidata's upper-level ontology indicating primary communities (top). Semantic network visualization of those communities (bottom).

Conclusion

The approach taken in this analysis serves as a starting point for studying the knowledge architecture, language, and prejudices of Wikidata. Employing ontology and semantic network analysis allows for examining how information is systematized in

Wikidata and how it mirrors societal trends. This study contributes to research on data infrastructures (Gray et al., 2018), the platformization of semantics (Iliadis et al., 2023), and the consequences of representing knowledge in organized forms in digital contexts. Beyond Wikidata, this research calls for radical transparency and criticism of proprietary AI knowledge systems to show their impact on society by allowing researchers to examine the classification architecture of databases used in consumer products.

References

Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *2008*. https://doi.org/10.1088/1742-5468/2008/10/P10008

Figueres-Esteban, M., Hughes, P., & Van Gulijk, C. (2016). Ontology network analysis for safety learning in the railway domain. In *Proceedings of the 26th European Safety and Reliability Conference, ESREL 2016*. CRC Press. https://eprints.hud.ac.uk/id/eprint/28633/

Ford, H. (2022). *Writing the revolution: Wikipedia and the survival of facts in the digital age*. MIT Press.

Ford, H., & Iliadis, A. (2023). Wikidata as semantic infrastructure: Knowledge representation, data labor, and truth in a more-than-technical project. *Social Media + Society*, *9*(3). https://doi.org/10.1177/20563051231195552

Fruchterman, T. M., & Reingold, E. M. (1991). Graph drawing by force-directed placement. *Software: Practice and Experience*, *21*(11), 1129-1164. https://doi.org/10.1002/spe.4380211102

Graham, M., Straumann, R. K., & Hogan, B. (2015). Digital divisions of labor and informational magnetism: Mapping participation in Wikipedia. *Annals of the Association of American Geographers*, *105*(6), 1158-1178. https://doi.org/10.1080/00045608.2015.1072791

Gray, J., Gerlitz, C., & Bounegru, L. (2018). Data infrastructure literacy. *Big Data & Society*, *5*(2). https://doi.org/10.1177/2053951718786316

Guzman, A. L., & Lewis, S. C. (2020). Artificial intelligence and communication: A Human–Machine Communication research agenda. *New Media & Society*, *22*(1), 70-86. https://doi.org/10.1177/1461444819858691

Hu, Y. (2005). Efficient, high-quality force-directed graph drawing. *Mathematica Journal*, *10*(1), 37-71.

Iliadis, A. (2022). *Semantic media: Mapping meaning on the internet*. Polity.

Iliadis, A., & Ford, H. (2023). Fast facts: Platforms from personalization to centralization. *Social Media + Society*, *9*(3). https://doi.org/10.1177/20563051231195546

Iliadis, A., Acker, A., Stevens, W., & Kavakli, B. (2023). One schema to rule them all: How Schema.org models the world of search. *Journal of the Association for Information Science and Technology*, *76*(2), 460-523. https://doi.org/10.1002/asi.24744

Koutsiana, E., Reklos, I., Saad Alghamdi, K., Jain, N., Meroño-Peñuela, A., & Simperl, E. (2024). Talking Wikidata: Communication patterns and their impact on community engagement in collaborative knowledge graphs. *arXiv preprint*. https://doi.org/10.48550/arXiv.2407.18278

Lurie, E., & Mustafaraj, E. (2018). Investigating the effects of Google's search engine result page in evaluating the credibility of online news sources. In H. Akkermans, K. Fontaine, I. Vermeulen, G-J. Houben, & M. S. Weber (Eds.), *WebSci '18: Proceedings of the 10th ACM Conference on Web Science* (pp. 107-116). https://doi.org/10.1145/3201064.3201095

Masullo, G. M., Wilhelm, C., Lee, T., Gonçalves, J., Riedl, M. J., & Stroud, N. J. (2024). Signaling news outlet trust in a Google Knowledge Panel: A conjoint experiment in Brazil, Germany, and the United States. *New Media & Society*, *26*(9), 5379-5402. https://doi.org/10.1177/14614448221135860

Nagel, T. (1986). *The view from nowhere*. Oxford. Oxford University Press.

Ray, L. (2020, March 2). 2020 Google search survey: How much do users trust their search results? *Moz*. https://moz.com/blog/2020-google-search-survey

Rothschild, A., Lurie, E., & Mustafaraj, E. (2019). How the interplay of Google and Wikipedia affects perceptions of online news sources. *Presented at the Computation + Journalism Symposium, Feb 2019, Miami, FL, USA*. https://emmalurie.github.io/docs/cplusj2019-interplay.pdf

Santos, V., Schwabe, D., & Lifschitz, S. (2024). Can you trust Wikidata?. *Semantic Web*. Preprint. https://www.semantic-web-journal.net/content/can-you-trust-wikidata-1

Segev, E. (Ed.). (2022). *Semantic network analysis in social sciences*. Routledge.

Shenoy, K., Ilievski, F., Garijo, D., Schwabe, D., & Szekely, P. (2022). A study of the quality of Wikidata. *Journal of Web Semantics*, *72*. https://doi.org/10.1016/j.websem.2021.100679

Vrandečić, D., & Krötzsch, M. (2014). Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, *57*(10), 78-85. https://doi.org/10.1145/2629489

Vrandečić, D., Pintscher, L., & Krötzsch, M. (2023). Wikidata: The making of. In Y. Ding, J. Tang, J. Sequeda, L. Aroyo, C. Castillo, G-J. Houben (Eds.), *WWW '23 Companion: Companion Proceedings of the ACM Web Conference 2023* (pp. 615-624). https://doi.org/10.1145/3543873.358557

Weng, S. S., & Chang, H. L. (2008). Using ontology network analysis for research document recommendation. *Expert Systems with Applications*, *34*(3), 1857-1869.

https://doi.org/10.1016/j.eswa.2007.02.023