University of Essex Online

MSc Artificial Intelligence

Report

Case Study Review

Submission date: 16.06.2025

Contents

Introduction to ITO

The landscape of Artificial Intelligence (AI) research has evolved into an increasingly complex ecosystem of models, benchmarks, tasks and evaluation methods. This diversity has highlighted the need for a structured and consistent framework capable of organizing and connecting all the complex components that define AI research and advancements (Blagec et al., 2022; Hogan et al., 2020). The Intelligence Task Ontology (ITO) was developed as a response to this need, providing a curated, extensible, and semantically rich infrastructure for describing and analyzing AI research outputs.

Unlike conventional platforms that catalog papers and performance scores, such as Papers with Code or State of the Art AI, the ITO framework provides a formal ontological structure that captures relationships between tasks, models, datasets, and evaluation metrics (Blagec et al., 2022; Martínez-Plumed et al., 2020). This enables researchers, students, developers, and policymakers to move beyond fragmented reporting and toward a more integrated understanding of how AI capabilities evolve across domains.

This report aims to demonstrate how formal ontological modelling, specifically with the ITO framework, can address challenges in knowledge representation, reasoning and benchmarking within AI research.

Summary of the ITO Framework

Background and Motivation

As outlined in the previous section Intelligence Task Ontology (ITO) was created to address the growing complexity in AI research. This was especially necessary as the number of publications rose from around 10’000 in 2000 to way more than 120’000 in 2019. This massive growth required a system for organization. Especially benchmarks and performance metrics needed to be aligned (Blagec et al., 2022; Zhang et al., 2021). Repositories such as Papers with Code may be valuable, but they offer limited semantic structure and they lack the ability to systematically interrelate AI tasks, input and output modalities, and evaluation strategies (Martínez-Plumed et al., 2020).

ITO was introduced to address these limitations by introducing a manually curated, ontology-based knowledge graph. It aims at facilitating a manually curated, ontology-based knowledge graph built on Semantic Web technologies including the Resource Description Framework (RDF) and the Web Ontology Language (WOL), allowing for reasoning and integration with external data sources. Its key strengths include a sophisticated semantic expressivity, interoperability, extensibility to other domains, support for automated inference, and facilitation of collaborative curation and meta-research.

Critical Evaluation

Strengths of ITO

The Intelligence Task Ontology (ITO) offers several significant strengths that distinguish it from existing AI benchmarking platforms. First, its foundation is based on formal Semantic Web standards. They used the earlier mentioned RDF and OWL which enables expressive, machine-interpretable representations of AI tasks, models, datasets, and performance metrics. This semantic infrastructure allows for reasoning, automated consistency checking, and integration with external ontologies such as EDAM from bioinformatics, Open Biomedical Ontologies (OBO), and Friend of a Friend (FOAF), fostering interoperability across domains (Blagec et al., 2022; W3C, 2012).

Furthermore, ITO places strong emphasis on manual curation, which enhances data accuracy and conceptual clarity. Its development involved several months of collaborative curation using tools like WebProtégé to normalize thousands of metric variations and align benchmark data to a polyhierarchical class structure (Blagec et al., 2022). This meticulous approach ensures high data integrity, setting ITO apart from more automated repositories that may suffer from inconsistency or lack of contextual annotation (Martínez-Plumed et al., 2020).

Additionally, ITO supports flexible and performant querying, particularly through the use of high-performance RDF graph databases like Blazegraph and Python libraries such as Owlready2. These tools allow users to run complex SPARQL queries across the ontology efficiently, even on standard computing hardware making it a great

choice for researchers (Blagec et al., 2022). This functionality is critical for large-scale meta-research applications, including studies on metric prevalence and longterm analysis of AI capability trends.

Finally, ITO has demonstrated clear utility in meta-research and domain-specific annotation. It has already been applied to the biomedical domain for cataloging over 450 datasets, and it simplifies literature reviews and task classification through its process-centric modeling (Blagec et al., 2021; Blagec et al., 2020). These features position ITO as a robust infrastructure for researchers and practitioners aiming to systematically evaluate progress in AI. This demonstrates ITO's value not only as a research aid but as a catalyst for more rigorous and transparent scientific inquiry in AI.

Weaknesses and Limitations

Despite its strengths, the Intelligence Task Ontology (ITO) also faces several notable limitations. One of the primary challenges is the complexity of the ontology itself, which can pose a steep learning curve for new users unfamiliar with semantic web technologies or ontological modeling. While tools like Protégé and BioPortal offer graphical interfaces, effectively navigating and querying a multi-thousand-class ontology still demands significant technical literacy (Blagec et al., 2022; Horridge et al., 2019). This creates a significant entry barrier for users, particularly those without prior experience in ontology engineering or Semantic Web tools.

A second limitation is ITO’s reliance on extensive manual curation. While this ensures high data quality, it also makes the ontology resource-intensive to maintain and scale, particularly as the AI landscape continues to evolve rapidly. Without sufficient automation or community-driven curation processes, the ontology may lag behind emerging tasks and benchmarks (Blagec et al., 2022). This limitation is especially relevant given the project's goal of periodic updates aligned with data from platforms like Papers with Code (PWC). This highlights that, despite its many strengths, the ITO framework also presents practical limitations that must be addressed to ensure its long-term sustainability and relevance.

Furthermore, certain design trade-offs such as the use of multiple inheritance introduce complexity into the class hierarchy. For example, the modeling decision to classify all benchmarks under both their specific AI task and a general benchmarking superclass creates a deeply nested and multi-parent structure. While this approach improves query performance, it deviates from best practices in ontology design and can hinder modularity and reuse (Blagec et al., 2022; Poveda-Villalón et al., 2014). This reflects a fundamental trade-off between performance optimization and the goal of following clean, modular ontology principles.

Finally, ITO’s dependence on PWC as its primary data source introduces a potential centralization risk. Although PWC is comprehensive, its data is semi-structured and partially crowd-sourced, meaning that any bias, inconsistency, or limitation in PWC is inherited by ITO. Alternative sources like AI Collaboratory or State of the Art AI were considered but ultimately found to overlap significantly with PWC or lacked open access (Blagec et al., 2022; Martínez-Plumed et al., 2020). This reliance raises

concerns about the robustness and representativeness of the knowledge graph in the long term.

Applications in Real-World

The Intelligence Task Ontology (ITO) offers various applications across AI research, industry benchmarking, and interdisciplinary collaboration. In research and development, ITO enhances transparency by providing a standardized and semantically grounded framework for representing and tracking AI model progress over time. Its structured performance metrics and task hierarchies enable meta-research on capability trajectories, such as identifying which subdomains of AI are advancing rapidly or stagnating (Blagec et al., 2022). This helps to take more informed research decisions and perform systematic mapping of trends across modalities and disciplines (Blagec et al., 2020).

In industry and benchmarking, ITO supports the selection of optimal models for specific tasks or domains by linking benchmark results to task-specific requirements and data modalities. Its ontological design helps stakeholders standardize performance evaluation, reducing inconsistencies in metric definitions and improving reproducibility. This is particularly valuable in regulated or high-stakes environments such as finance, healthcare, or autonomous systems (Blagec et al., 2022). Developers and researchers can query ITO to compare models not just by score, but by task complexity, input-output formats, or data type compatibility.

Looking toward collaboration and future potential, ITO enables cross-domain integration by linking with external ontologies and knowledge graphs such as the Artificial Intelligence Knowledge Graph (AI-KG), the Open Research Knowledge Graph (ORKG), and the Computer Science Ontology (CSO). This facilitates interdisciplinary projects in fields such as medicine, robotics, and cognitive science,

where AI tasks intersect with rich domain-specific knowledge (Dessì et al., 2020; Jaradeh et al., 2019). Its modular structure and use of Semantic Web standards make it a strong candidate for ontology enrichment, allowing ongoing integration of new AI domains, tasks, and benchmarks as the field evolves.

Conclusion

The Intelligence Task Ontology (ITO) is a foundational initiative that brings structure to the fragmented AI research landscape. By combining manual curation with Semantic Web technologies, it offers a robust framework for organizing tasks, benchmarks, and metrics. ITO supports meta-research, reproducibility, and informed evaluation.

As a bridge across disparate AI data sources, ITO connects models, datasets, and performance metrics in a unified ontology. Its relevance spans academia and industry, enabling better benchmarking and knowledge integration.

Future improvements should focus on automating data updates, enhancing accessibility, and linking with other knowledge graphs. Through continued collaboration, ITO can remain a vital tool for transparent and coordinated AI development.

References

Blagec, K., Barbosa-Silva, A., Ott, S. and Samwald, M., 2022. A curated, ontology-based, large-scale knowledge graph of artificial intelligence tasks and benchmarks. *Scientific Data*, *9*(1), p.322.

Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., Melo, G.D., Gutierrez, C., Kirrane, S., Gayo, J.E.L., Navigli, R., Neumaier, S. and Ngomo, A.C.N., 2021. Knowledge graphs. *ACM Computing Surveys (Csur)*, *54*(4), pp.1-37.

MARTINEZ, P.F., HERNÁNDEZ-ORALLO, J. and GOMEZ, G.E., Tracking AI: The Capability Is (Not) Near. *FRONTIERS IN ARTIFICIAL INTELLIGENCE AND APPLICATIONS*.

Blagec, K., Kraiger, J. and Samwald, M., 2021. A living catalogue of artificial intelligence datasets and benchmarks for medical decision making. *Zenodo https://doi. org/10.5281/zenodo*, *4647824*.

Blagec, K., Dorffner, G., Moradi, M. and Samwald, M., 2020. A critical analysis of metrics used for measuring progress in artificial intelligence. *arXiv preprint arXiv:2008.02577*.

W3C, 2012. *OWL 2 Web Ontology Language Document Overview (Second Edition)*. [online] Available at: https://www.w3.org/TR/owl2-overview/

Horridge, M., Gonçalves, R.S., Nyulas, C.I., Tudorache, T. and Musen, M.A., 2019, May. Webprotégé: A cloud-based ontology editor. In *Companion Proceedings of The 2019 World Wide Web Conference* (pp. 686-689).

Poveda-Villalón, M., Gómez-Pérez, A. and Suárez-Figueroa, M.C., 2014. Oops!(ontology pitfall scanner!): An on-line tool for ontology evaluation. *International Journal on Semantic Web and Information Systems (IJSWIS)*, *10*(2), pp.7-34.

Dessì, D., Osborne, F., Reforgiato Recupero, D., Buscaldi, D., Motta, E. and Sack, H., 2020. AI-KG: an automatically generated knowledge graph of artificial intelligence. In *The Semantic Web–ISWC 2020: 19th International Semantic Web Conference, Athens, Greece, November 2–6, 2020, Proceedings, Part II 19* (pp. 127-143). Springer International Publishing.

Jaradeh, M.Y., Oelen, A., Farfar, K.E., Prinz, M., D'Souza, J., Kismihók, G., Stocker, M. and Auer, S., 2019, September. Open research knowledge graph: next generation infrastructure for semantic scholarly knowledge. In *Proceedings of the 10th international conference on knowledge capture* (pp. 243-246).