Analysis and Application of the Intelligence Task Ontology (ITO) in AI Benchmarking

University of Essex, MSc in AI, Module 4

Pavlos Papachristos, PP24589

Contents

Analysis and Application of the Intelligence Task Ontology (ITO) in AI Benchmarking

1. Background and Summary of the ITO Framework

The Intelligence Task Ontology (ITO) is developed to tackle the rapid growth of AI models and benchmarks which has introduced an increasing fragmentation and inconsistency in AI benchmarking, on the tracking of the task evolution and on the models performance comparisons.

To address these challenges, the ITO methodology is developed with the use of the OWL 2 and the Resource Description Framework (RDF) to represent multiple AI tasks, models, benchmarks, and performance metrics (Blagec et al., 2022). The ITO is designed and built as a machine-readable semantic structure.

The OWL 2 (Web Ontology Language 2) is a W3C standard for creating, sharing, and reasoning over ontologies. That is a set of formal representations of knowledge with well-defined semantics. This is an enhancement of the original OWL 1 (2004), released as a recommendation by W3C in 2009, and provides more expressive power, better tool support, and richer modelling features.

The OWL 2 is built on top of the RDF and uses Description Logics (DLs) as its formal foundation — enabling reasoning, classification, and logical inference over structured data. The DLs is a group of formal knowledge representation languages that are designed to describe and reason on the conceptual structure of a domain in a precise and computable way.

The RDF is a W3C standard for representing information about resources on the web. The RDF forms the foundation of the Semantic Web that allows data to be structured in a way that is machine-readable and semantically rich.

The Semantic Web is a form of the web where the information is given with a well-defined meaning which enables computers and people to work together more intelligently. It was proposed by Berners-Lee et. al. (2001) as an extension of the current web (Web 2.0) where machines can process text, but they don’t understand its meaning. The Semantic Web introduces standards that allow for automated reasoning and intelligent services.

The ITO project aimed to align with FAIR principles to promote open and reusable scientific data. ITO incorporates a large number of RDF triples (the RDF triple is a 3-part statement that describes a relationship between resources) covering more than 1,100 tasks and 3,600 datasets, offering a comprehensive semantic infrastructure for AI benchmarking.

The FAIR principles — Findable, Accessible, Interoperable, and Reusable — can be implemented directly in OWL ontologies through the use of semantic web standards, structured metadata, and best practices in ontology engineering. OWL (Web Ontology Language), being a W3C standard built on RDF and Description Logic, is natively designed to support FAIR data management.

2. ITO Framework: Goals, Methodology and Findings

The main objectives of ITO are to: (a) enable structured representation of AI knowledge, (b) support FAIR data integration, (c) harmonize benchmark standards, and (d) support reasoning and semantic queries through open knowledge graph structures.

The methodology applied uses ITO to integrate data from Papers with Code using RDF triples. With the manual curation the harmonisation of naming inconsistencies and taxonomic inaccuracies is achieved. Tools such as Protégé and HermiT were used for ontology building and reasoning. The core entities include Tasks, Models, Datasets and Metrics and linked with properties like 'evaluatedOn', 'hasInput', and 'achievesMetric'.

The resulting ontology comprises 685,000 RDF triples, representing 1,100+ tasks and 3,600+ datasets. The implementation of the ITO framework helps to expose semantic gaps in current AI benchmarks, facilitates inference and supports polyhierarchical classification of AI domains.

The multimodal nature of modern AI tasks (e.g., visual question answering under both NLP and computer vision) accommodated with the use of Polyhierarchical task classification. To facilitate the comparison and interpretability between models, the performance metrics were normalized and classified within a semantic taxonomy.

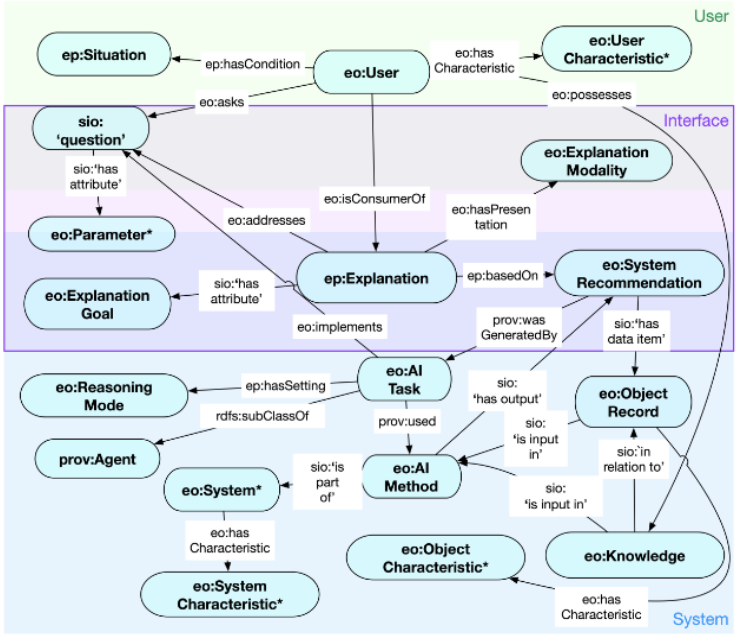
The ITO's methodology facilitates the integration of OWL and RDF for formal semantics, allowing logical inference and structured benchmarking. Its polyhierarchical design supports complex task classification across AI modalities.

Using manual curation the accuracy is improved but this has resulted in the introduction of scalability limits. The use of alternative methods such as knowledge embedding graphs or pretrained LLM-based classification offers a broader scalability but they lack semantic precision.

The major strength of the ITO framework is its extensibility and support for logical inference, which outperforms flat data repositories and spreadsheets.

Human annotation ensured accuracy but may not keep pace with fast-moving benchmarks. Compared to flat taxonomies or label-based systems, ITO excels in expressivity, support for open science and alignment with explainability standards advocated by Confalonieri and Guizzardi (2022).

In Figure 1 the overview of the Ontology-Based Explanation Framework is outlined:



(source: Knowledge Representation & Reasoning, notes Week 6, Sammy Danso, University of Essex, June 2025).

3. Assessment of the ITO Framework

One of the major drawbacks of the ITO lays with its reliance on manual curation which introduces scalability issues. In particular, over 60 metric variants were harmonized manually which is a labor-intensive process. This improves the semantic accuracy, but it also slows down the updates.

Alternative approaches like knowledge embeddings or NLP-driven auto-tagging may scale better but lack the explainability and structure provided by ITO's ontological design. The user-interface-system explanation framework (Confalonieri & Guizzardi, 2021) further demonstrates how ontologies enhance interpretability by aligning explanations with human conceptual models and common-sense reasoning.

In their article Paton and Kobayashi (2023) emphasise in the importance of open, modular, and reproducible AI development in the healthcare sector, incorporating FAIR data standards and open-source tooling. Their approach bridges gaps between clinical practitioners and AI researchers, improving interpretability, reliability, and ethical alignment in health AI applications. By applying these practices, models developed in healthcare can better adapt to shifting patient populations and clinical settings.

Confalonieri and Guizzardi examine how ontologies serve not only as background knowledge repositories but also as structural scaffolds for Explainable AI (XAI). The

application of the Ontologies help to enhance interpretability by linking data to domain concepts, improving causal explanations and supporting argumentative reasoning.

In particular, the authors argue that ontology-driven XAI allows better alignment with human cognition and domain expert understanding. The multiple conceptual, methodological and practical roles that the Ontologies play within XAI systems were outlined.

4. Real-World Applications and Implications

The ITO methodology aligns with open science practices by promoting data interoperability and accessibility. As a result it can enhance the AI model governance and serve as a foundation for explainable AI pipelines. Furthermore, ITO may be integrated into research data repositories, used in healthcare AI model validation (as per Paton & Kobayashi, 2022) and adopted in academic publishing for performance traceability. Its extensibility and FAIR compliance make it a strong candidate for long-term AI evaluation infrastructure.

In the real-world applications the implementation of the ITO framework can transform the AI research practices and standardization by:

- Enhancing the AI model governance with the use of traceable and reproducible benchmarks.

- Offering a reliable methodology for the standardisation of task definitions and references and improving the model benchmarking, which can support the academic research and publications.

- Integrate with decision-tree explanations in regulated domains like finance or healthcare, enhancing transparency and compliance.

- Improving the interpretation of the AI pipelines with mapping the model performance to understandable semantic layers.

- Facilitating meta-analysis and automating audits in large-scale research repositories.

In healthcare, for example, ITO's integration of FAIR principles and semantic traceability aligns well with the open science principles proposed by Paton & Kobayashi (2023).

The incorporation of the ITO framework into model validation tools could assist to assess the AI tool relevance and performance for specific diagnostic tasks. The framework is also compatible with explainability needs, such as those illustrated by decision trees and layered ontology systems (as shown in Confalonieri & Guizzardi, 2021).

4.1 Use Case: Financial Model Risk Management - Inventory and Validation Governance

The ITO framework presents a great potential for applications in financial services and in particular in the area of model validation, risk governance and regulatory compliance.

Under regulatory frameworks such as the Federal Reserve's SR 11-7 and the UK's PRA SS1/23, banks are mandated to maintain robust model risk management practices.

ITO supports this by providing a formal, query enabling framework for defining models by category (e.g., IRRBB, ALM, fraud detection), validation status, performance metrics, and reviewer audit trails. Ontology-based tagging enables enhanced control and compliance, offering linked metadata for each model instance.

ITO can integrate with explainability ontologies (e.g., Confalonieri & Guizzardi, 2021) to support regulatory expectations for transparency and control over AI or Algorithmic trading models that frequently characterised as black-boxes.

Banks can configure rules—such as flagging all such models with high-materiality that require special governance. Through OWL-based inference, model classification and risk-tiering can be dynamically maintained.

The possibility of using the ITO's SPARQL-based querying can automate parts of regulatory reporting. The SPARQL (SPARQL Protocol and RDF Query Language) is the official W3C query language used to retrieve and manipulate data stored in RDF format. It is the semantic web's equivalent of SQL for relational databases — but instead of querying tables, SPARQL queries "triples" in RDF graphs.

For example, identifying models lacking benchmark documentation, unvalidated models beyond acceptable timeframes, or models using deprecated datasets becomes possible with semantic queries. This functionality aligns well with SS1/23's emphasis on model lifecycle control and governance traceability.

Additionally, using RDF's triple structure, the ITO framework enables the representation of task–model–dataset–metric relationships that facilitates the model reuse and their consistent deployment across business lines which is aligned with the FAIR data principles.

Benefits

The implementation of ITO in Financial Model Risk Management Governance and Control, enables a consistent and auditable model classification system. It enhances the explainability and governance for the AI and algorithmic trading models and automates the model lifecycle reporting in compliance with the regulatory requirements.

The ITO framework is aligned with the FAIR and open science principles and it is fostering the model reuseability, discoverability and traceability.

Challenges and Considerations

While the application of ITO in finance is promising, it does require financial institutions to adopt existing infrastructures which requires investment to strategic solutions that are costly and may take to materialise.

Efforts such as ontology alignment with financial tasks, integration with legacy model governance platforms, and staff upskilling in semantic technologies (e.g., Protégé, OWL, SPARQL) are necessary for successful implementation.

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. https://doi.org/10.1038/s41597-022-01435-x
- Berners-Lee, T., Hendler, J., & Lassila, O., 2001. The semantic web. Scientific American, 284(5), pp.28–37. https://doi.org/10.1038/scientificamerican0501-34
- Confalonieri, R. and Guizzardi, G., 2021. On the multiple roles of ontology in explainable AI. Applied Ontology, 16(2), pp.141–167. https://doi.org/10.3233/AO-210255
- Paton, C. and Kobayashi, S., 2023. An open source approach to AI in healthcare. BMJ Health & Care Informatics, 30(1). https://doi.org/10.1136/bmjhci-2023-100783
- PRA, 2023. Supervisory Statement SS1/23: Model risk management principles for banks. Bank of England. https://www.bankofengland.co.uk/prudential-regulation/publication/2023/june/model-risk-management-principles-for-banks-ss
- SR 11-7, 2011. Supervisory Guidance on Model Risk Management. Board of Governors of the Federal Reserve System. https://www.federalreserve.gov/supervisionreg/srletters/sr1107.htm
- Wilkinson, M.D. et al., 2016. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3, p.160018. https://doi.org/10.1038/sdata.2016.18
- W3C, 2024. Semantic Web. World Wide Web Consortium. https://www.w3.org/2001/sw/