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

Abstract

Tracking progress in artificial-intelligence (AI) research has become difficult as conference papers quadrupled and journal articles climbed from ≈10 000 from 2000 to >120 000 in 2019 (Zhang et al., 2021, cited in Blagec et al., 2022). The Intelligence Task Ontology and Knowledge Graph (ITO) addresses this challenge by storing AI tasks, benchmarks, and metrics in a FAIR, machine-readable graph of 1,100 task classes and 1,995 metric properties (Blagec et al., 2022). This case study summarises ITO, critiques its coverage, design, and sustainability, and explores four practical use-cases ranging from research-policy, real world industrial applications and regulatory audits thereby assessing whether formal knowledge-representation can support evidence-based AI development and benchmarking

1. Introduction

Explosive growth in AI literature fragments knowledge, hampers reproducibility and complicates evidence-based decisions (Stanford HAI, 2023). Formal knowledge-representation and reasoning (KR&R) tools such as ontologies, description logics, rule systems offer a remedy. ITO operationalises those tools by embedding a curated hierarchy of AI processes and metrics in OWL 2, enabling SPARQL queries and automated reasoning. This study evaluates how well ITO mitigates benchmark saturation (Ott et al., 2022) and metric proliferation (Sambasivan et al., 2021).

2. Case Study Summary

2.1 Background and Motivation

The AI Index shows journal outputs grew tenfold in under two decades, while arXiv uploads rose six-fold in five years (Zhang et al., 2021). Task names and metric labels vary widely, and leaderboard sites age quickly. To impose order, Blagec et al. built **ITO**, guided by five requirements: (i) manual curation of hierarchies; (ii) graph representation; (iii) easy external linking; (iv) automated reasoning and validation; and (v) ongoing community editing via WebProtégé (Blagec et al., 2022).

2.2. Methodology

ITO was constructed through a semi-automated pipeline that turns raw Papers-with-Code data into a continuously updated ontology release (Blagec *et al.*, 2022):

- **Data Source** - All benchmark records were taken from the Papers-with-Code (PWC) repository, which listed >5 000 benchmarks and ~50 000 papers at the time.
- **Automatic import** – A custom Python loader translated the raw PWC JSON dump into RDF/OWL triples that form the backbone of the knowledge graph (Blagec *et al.*, 2022).
- **Manual curation** – Two AI experts cleaned task and metric labels in WebProtégé, reused terms from EDAM, OBO, Dublin Core and FOAF, and organised tasks under **16 top-level "AI process" classes** such as *Natural-language processing* and *Vision process*.
- **Metric clean-up** – More than **800 raw metric strings** were collapsed into a canonical hierarchy of **1 995 metric properties**, turning dozens of spelling variants of the same score into one standard term.
- **Incremental updates** – New items imported by script are marked *"requiring curation"* so curators can approve them before each release, keeping the graph current without losing quality (Blagec *et al.*, 2022).

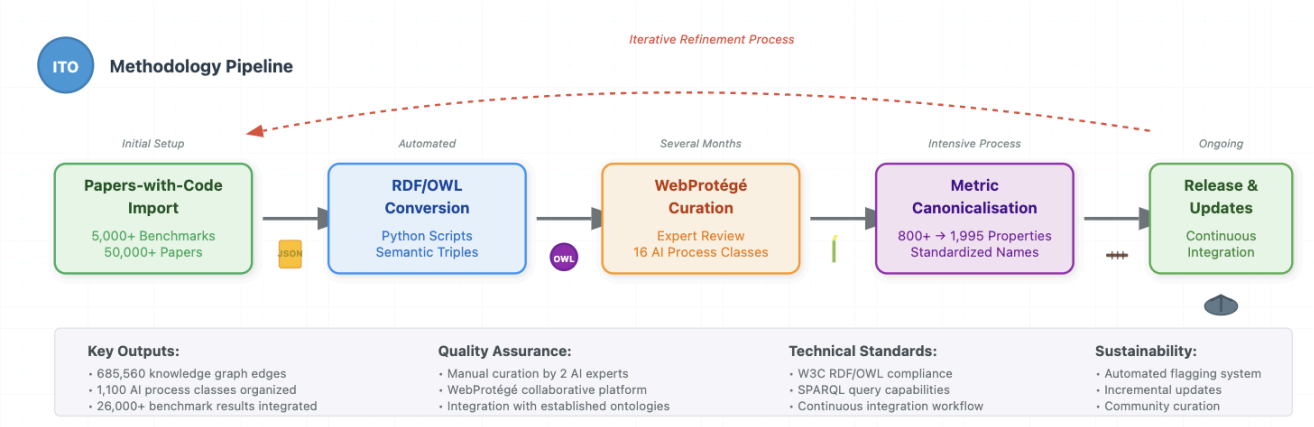

Figure 1: ITO Development Pipeline: Semi-automated workflow transforming Papers-with-Code data into a continuously updated ontology through expert curation and standardization processes.

2.3 Key findings (ITO v1.01)

- **Scale** – The graph contains 685 560 edges, 1,100 AI-process classes, >50 000 individuals and 26 000 benchmark results covering 3 633 datasets from 2000-2021 (Blagec *et al.*, 2022).

- **Task coverage** – Computer-vision and NLP processes hold the most benchmarks, while areas like audio and graph processing have fewer, revealing under-explored fields.

- **Metric standardisation** – Canonical names replace 60+ spelling variants per metric in some cases, allowing fair cross-paper comparisons.

- **Meta-research potential** – Because every result is timestamped and typed, users can trace progress curves, spot benchmark saturation and study links between tasks, datasets and methods over time (Blagec *et al.*, 2022).

3. Critical Evaluation

3.1 Ontology design & representation

ITO adopts a process-centric poly-hierarchy in OWL 2, giving each task class logical parents such as Natural-language processing and Vision process (Blagec et al., 2022).

- **Strengths**: Deep class depth (average ≈ 5.3) and a high axiom-to-class ratio (>9) enable description-logic (DL) reasoning and precise SPARQL queries. (Blagec et al., 2022, Table 4)

- **Limitations**: The single superclass Benchmarking yields an extreme sibling count (>4 500) and high tangledness, contravening modular-design advice in ontology engineering manuals (Raad & Cruz, 2015). Multiple inheritance also complicates maintenance once finer orthogonal facets (e.g. modality, objective) are layered in.

3.2 Coverage, completeness and bias

Because all seed data come from Papers-with-Code (PWC), ITO mirrors that platform's focus: computer vision and NLP host most of the 26, 000 benchmark results, whereas audio and graph tasks are sparse (Blagec et al., 2022). Roughly 49 % of classes lack instances helpful scaffolding, but a signal that empirical coverage is uneven. Similar domain skew has been reported for AI-KG, which over-represents deep-learning topics mined from arXiv (Dessì et al., 2020).

3.3 Metric handling and interoperability

ITO compresses **>800 free-text metric labels into 1 995 canonical properties**, removing spelling noise (Blagec *et al.*, 2022).

- **Positive:** Canonical URIs permit longitudinal and cross-task analytics absent from most public leaderboards.

- **Issues:** Formulas for some metrics (e.g. *Top-k accuracy*) are not always stored, echoing the ambiguity documented by Blagec *et al.* (2020) in their survey of AI-metric misuse. Moreover, most industrial pipelines track metrics in JSON Schema or MLflow; adapters are still required, as noted in ORKG integration studies (Jaradeh *et al.*, 2019).

3.4 Update workflow and sustainability

The import script tags new items *Meta: requiring curation*, letting humans vet changes before release. Yet the v1.01 team consisted of **only two curators** (Blagec *et al.*, 2022); scaling to PWC's projected 100 000+ papers will demand NLP-assisted normalisation or a larger crowd. Community-governance models such as those used in CSO (Salatino *et al.*, 2020) or Wikidata could serve as templates.

3.5 Comparative landscape

Criterion	ITO	AI-KG (Dessì *et al.*, 2020)	CSO (Salatino *et al.*, 2020)	State-of-the-Art AI (web)
Downloadable dump	**Yes (OWL/RDF)**	Yes (RDF)	Yes (OWL)	No
Manual curation	**Yes (tasks & metrics)**	No	No	Crowd tags
Performance results stored	**26 000**	None (paper sentences only)	None	≈ leader-board counts
DL-reasoner ready	**Yes**	Limited	Limited	No
Active maintenance (2025)	**Yes (v1.01)**	Yes	Yes	UI only

Table 1: Comparative landscape (Blagec et al., 2022; Dessì et al., 2020; Salatino et al., 2020)

3.6 Overall strengths and weaknesses

Strengths	Weaknesses
FAIR publication (Zenodo DOI, CC-BY-SA) and example Jupyter notebooks enhance reproducibility.	Single-source dependence on PWC risks perpetuating its topical bias.
SHACL validation plus Protégé-ELK checks keep the graph logically consistent.	Manual curation is labour-intensive; scaling without crowdsourcing or NLP aids is uncertain.
Timestamped triples enable meta-research on progress curves and benchmark saturation (cf. Ott *et al.*, 2022).	Design choice of a huge **Benchmarking** superclass breaks best-practice modularity, making refactoring harder.

Table 2: Overall strengths and weaknesses

4. Real-world applications and implications

The curated, machine-readable structure of ITO means it can be dropped into several practical workflows that currently rely on ad-hoc spreadsheets or fragile web scraping. Below are four concrete application areas, each tied to published or emerging practice.

4.1 Research policy and funding

Because every benchmark result in ITO is time-stamped and typed, national science agencies can query longitudinal progress curves instead of relying on anecdotal “state-of-the-art” claims. For example, plotting top-1 ImageNet accuracy from 2012-2024 shows a deceleration after 2021; funding bodies could redirect grants toward less-saturated tasks such as multimodal reasoning or graph learning. Early pilots using ITO for capability-trend analysis are reported in Barbosa-Silva et al. (manuscript in preparation, cited in Blagec et al., 2022). This aligns with the wider meta-research agenda that Ioannidis (2018) advocates using structured data about science to improve science itself. (Blagec et al., 2022; Ioannidis, 2018).

4.2 Retrieval-Augmented Generation (RAG) pipelines

Modern LLM pipelines combine vector search with symbolic graphs for grounding. Plugging ITO into that layer allows models to answer, for example, “Which benchmarks test commonsense reasoning?” while citing the exact dataset and metric URIs. Enterprise studies show domain graphs can cut hallucination rates threefold (data.world Team, 2023; Derbier, 2025; Blagec *et al.*, 2022).

4.3 Benchmark governance and reproducibility

Logging ITO URIs in MLflow (a platform for developing Machine Learning models) enables automatic checks that, for example, *BLEU* scores follow the *sacreBLEU* protocol, addressing reproducibility gaps highlighted by ML conference meta-reviews. (Blagec *et al.*, 2020).

4.4 Regulatory compliance & Industry standardisation

Regulators are moving towards mandatory reporting of AI system capabilities (EU AI Act, 2024 draft). A shared ontology of tasks and metrics fulfills the "common specification" clause, easing compliance audits. Early pilots in healthcare AI benchmarking already leverage domain-specific knowledge graphs (Wall Street Journal, 2025).

Conclusion

The Intelligence Task Ontology and Knowledge Graph (ITO) demonstrates how formal knowledge-representation can tame the sprawl of modern AI benchmarking: its curated OWL classes, canonical metric hierarchy and FAIR publication create a reproducible backbone for meta-research, experiment tracking and regulatory audits. Yet our review shows that ITO's reliance on Papers-with-Code skews coverage toward vision and NLP, while manual curation and a monolithic Benchmarking superclass threaten maintainability and modularity. Comparisons with AI-KG and CSO confirm ITO's unique value including downloadable data plus rich semantics, but also underline the need for broader data sources, crowd-based governance and tighter integration with ML-ops tooling. Addressing these gaps will determine whether ITO matures into enduring infrastructure or remains a well-crafted snapshot of today's AI landscape.

References

Blagec, K., Barbosa-Silva, A., Ott, S. & Samwald, M. (2022) 'A curated, ontology-based, large-scale knowledge graph of artificial intelligence tasks and benchmarks', *Available online at https://www.nature.com/articles/s41597-022-01435-x#Sec5 [Accessed on 16th July 2025)*

Blagec, K., Dorffner, G., Moradi, M. & Samwald, M. (2020) *A critical analysis of metrics used for measuring progress in artificial intelligence*. arXiv:2008.02577. Available online via https://arxiv.org/pdf/2008.02577 *[Accessed on 16th July 2025)*

Dessì, D. *et al.* (2020) 'AI-KG: An automatically generated knowledge graph of artificial intelligence', Available online at https://oro.open.ac.uk/71736/1/_ISWC_2020___resources__AI_Knowledge_Graph%20%281%29.pdf *[Accessed on 16th July 2025)*

European Commission (2024) *Proposal for a Regulation laying down harmonised rules on artificial intelligence (AI Act)*. Brussels.

Ioannidis, J.P.A. (2018) 'Meta-research: Why research on research matters', *PLoS Biology*, 16(3), e2005468.

Jaradeh, M.Y. *et al.* (2019) 'Open research knowledge graph: next-generation infrastructure for semantic scholarly knowledge', in *K-CAP '19*.

Ott, S. *et al.* (2022) 'Mapping global dynamics of benchmark creation and saturation in AI', *Nature Communications*, 13, 6732.

Raad, J. & Cruz, C. (2015) 'A survey on ontology evaluation methods', in *IC3K 2015 Proceedings*,

Salatino, A.A. *et al.* (2020) 'The Computer Science Ontology', *Data Intelligence*, 2(3), 379-416.

Sambasivan, N. *et al.* (2021) 'Everyone wants to do the model work, not the data work', *CHI 2021 Proceedings*

Stanford HAI (2023) *AI Index Report 2023*. Stanford University. Available online at https://hai.stanford.edu/assets/files/hai_ai-index-report_2023.pdf [*Accessed on 16th July 2025]*

data.world Team (2023) 'Generative AI benchmark: Increasing the accuracy of LLMs in the enterprise with a knowledge graph', 13 Nov.

Wall Street Journal (2025) 'AI-powered databases boost the Alzheimer's drug-discovery process'. Available online at https://www.wsj.com/articles/ai-powered-databases-boost-the-alzheimers-drug-discovery-process-b9b75180 [*Accessed on 16th July 2025]*