

From Documents to Decisions: Ontology-Driven Knowledge Graph Construction Using Iterative LLM Pipelines

Felix Neubürger^{1,2}, Ole Katzenberger^{1,2}, Salvador Cantuarias² and Thomas Kopinski^{1,2}

¹South Westphalia University of Applied Sciences

²GAIA Lab

Abstract

Organisations in specialised domains—nuclear engineering, legal compliance, industrial asset management—make critical decisions based on knowledge scattered across heterogeneous document collections. We present an iterative, ontology-driven pipeline that transforms unstructured documents into structured, queryable knowledge graphs operating entirely on local large language models (LLMs). The system combines three advances: (1) competency-question (CQ) driven discovery loops that autonomously identify and fill knowledge gaps; (2) tiered extraction with a six-stage confidence pipeline producing ontology-consistent triples; and (3) integrated SHACL validation during construction. A six-condition ablation study on a nuclear decommissioning corpus (33 documents) shows that multi-pass iterative discovery increases entity count by +51.7% and relation coverage from 41.3% to 74.3%; law-graph augmentation further improves average node degree from 7.19 to 12.58 and raises SHACL conformance to 0.953. Domain transferability is demonstrated on 17 German federal statutes using identical pipeline code.

Keywords

Knowledge Graph Construction, Ontology-Driven Extraction, Large Language Models, Competency Questions, SHACL Validation

1. Introduction

In specialised domains such as nuclear engineering and legal compliance, critical decisions depend on synthesising knowledge from large, heterogeneous document collections [1, 2]. Knowledge graphs (KGs) transform scattered, unstructured information into structured, queryable representations that support automated reasoning and retrieval-augmented question answering. Constructing these graphs from mixed document formats remains labour-intensive and typically requires costly cloud-based LLM APIs or extensive manual curation.

Recent surveys confirm that LLM-based extraction struggles with ontological consistency: triples frequently violate domain constraints, miss required relations, or introduce hallucinated entities [3, 4]. Schema-free approaches achieve broad coverage but yield noisy graphs unsuitable for decision support [5]. Ontology-guided approaches improve precision but treat the ontology as a static schema—they cannot autonomously discover what the schema *does not yet capture*.

No existing end-to-end system combines: (1) ontology-driven discovery loops using competency questions; (2) confidence-calibrated tiered extraction; (3) SHACL validation integrated into construction; and (4) operation entirely on local LLMs for data-sensitive environments.

Contributions

1. **CQ-driven discovery loop** that autonomously identifies coverage gaps and iterates until convergence.
2. **Six-stage confidence pipeline** combining rule-based and LLM-based extraction with multi-stage calibration.

Posters, Demos, Blue Sky, and Tutorials at SEMANTiCS 2026, Sep 2026, Ghent, Belgium

✉ neubuerger.felix@fh-swf.de (F. Neubürger)

ORCID 0000-0002-8983-6466 (F. Neubürger)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

3. **Integrated SHACL validation** enforcing ontological consistency during construction, not post-hoc.
4. **Domain-agnostic architecture** demonstrated on nuclear decommissioning and German law—only ontology and corpus change.

The pipeline is iterative and goal-directed: CQs derived from the ontology define coverage targets, the system autonomously detects gaps after each pass, and new CQs are generated to close them—a loop that terminates when improvement falls below a threshold or maximum iterations are reached. Code and data are publicly available.¹

2. Related Work

LLM-based KG construction surveys distinguish schema-based (higher precision) from schema-free (higher recall) approaches [3, 4]. Ontology grounding improves consistency [6], and multi-agent decomposition can structure extraction [7], but existing systems lack autonomous CQ-driven discovery loops. LLM assistance for ontology engineering is maturing [9, 10], yet axiom correctness remains challenging. Our work integrates CQ-driven gap detection directly into the construction loop, combining rule-based and LLM extraction with SHACL feedback.

3. System Architecture

3.1. Overview

The system implements a three-layer architecture—*extraction*, *enrichment*, and *persistence*—wrapped in a CQ-driven discovery loop. All services (FastAPI orchestration, Ollama LLM backend, Neo4j 5.x, Qdrant, Apache Fuseki 4.x) run in a single Docker Compose stack on commodity hardware. The architecture is *ontology-agnostic*: switching domains requires only a new ontology and document corpus; all 12 pipeline phases remain identical. The system comprises 34.8k lines of Python with 889 test functions and processes six document formats (PDF, DOCX, PPTX, XML, HTML, CSV).

3.2. CQ-Driven Discovery Loop

The discovery loop (Figure 1) drives all 12 pipeline phases:

1. **CQ generation**: The OWL ontology is queried via SPARQL; class definitions seed competency questions (configurable per class).
2. **Hybrid retrieval**: Dense embedding search (4096-dim qwen3-embedding) plus BM25 sparse retrieval, followed by cross-encoder reranking, targets relevant document chunks per CQ.
3. **Tiered extraction**: A `RuleBasedExtractor` applies regex patterns and domain dictionaries first; an `LLMEntityExtractor` falls back when rules yield insufficient results. Structured output is enforced via a Pydantic schema with automatic retry.
4. **Relation extraction**: A second LLM pass extracts relations constrained by ontology property definitions (domain, range, cardinality).
5. **Confidence scoring and SHACL validation**: All extracted triples pass through a six-stage confidence pipeline (Section 3.3) and are validated against SHACL shapes before persistence.
6. **Coverage measurement and convergence**: The loop terminates when coverage improvement drops below 2%, CQ answerability exceeds 80%, or maximum iterations are reached; otherwise gap-targeted follow-up CQs are generated.

Layer 1 results are checkpointed; Layers 2–3 re-run from checkpoint in ~20 minutes—a 94% time saving enabling rapid iterative refinement.

¹<https://github.com/DataScienceLabFHSWF/KnowledgeGraphBuilder>; dataset: <https://doi.org/10.5281/zenodo.20068694>

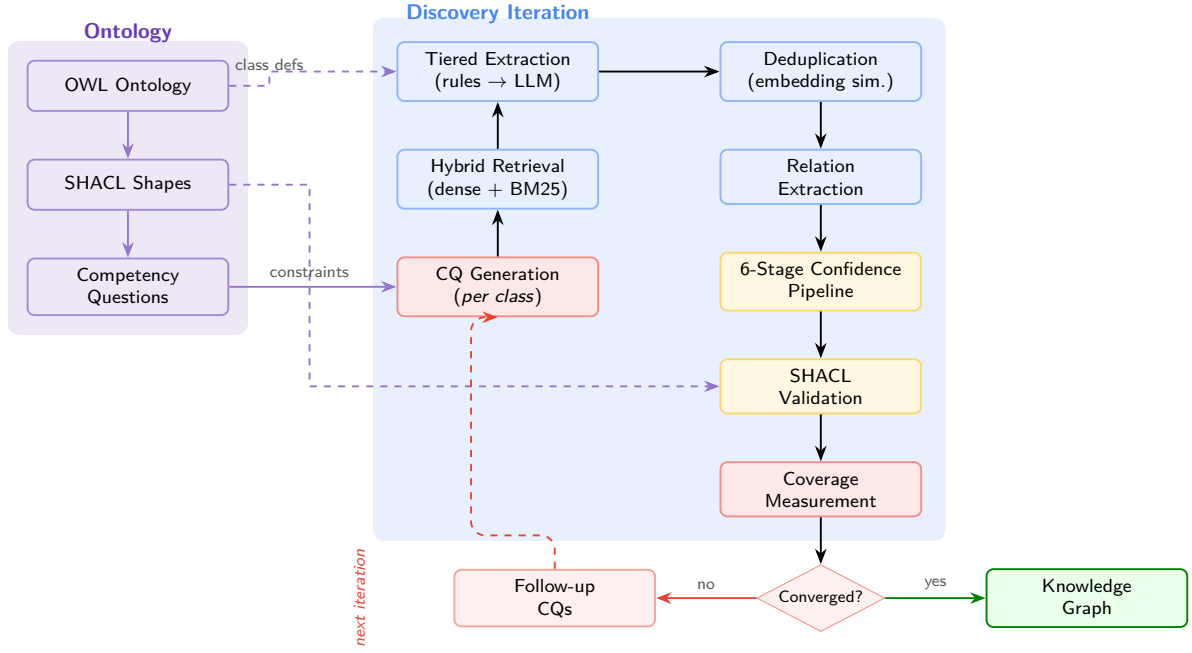


Figure 1: CQ-driven discovery loop. The OWL ontology (left) seeds class-specific competency questions. Each iteration follows the blue flow: hybrid retrieval, tiered extraction, deduplication, relation extraction, six-stage confidence scoring, and SHACL validation. Coverage measurement (bottom) drives a convergence decision; unmet targets feed back as gap-targeted CQs (dashed red arrow).

3.3. Six-Stage Confidence Pipeline

All extracted entities pass through a sequential confidence refinement pipeline before enrichment: (1) *Base scoring* from the extractor (rule-based: 0.90–1.0; LLM-based: calibrated per ontology class); (2) *Distributional analysis* comparing entity type frequencies against ontology class priors; (3) *Multi-source evidence boosting* when the same entity is found in multiple document chunks; (4) *Isotonic calibration* fitting a monotone regression to align scores with empirical acceptance rates; (5) *Consensus voting*: an EnsembleExtractor runs multiple extractors in parallel; entities confirmed by $k \geq 2$ extractors receive a confidence boost $c' = \min(0.99, \bar{c} + \min(0.25, (k - 1) \times 0.1))$; (6) *Quality filtering*: entities below threshold (default 0.60) are removed. SHACL violation patterns feed back into subsequent iterations as targeted refinement signals.

3.4. Law Graph and Cross-Domain Linking

For domains intersecting regulatory frameworks, the pipeline constructs a parallel *law graph* from 17 German federal statutes (XML, *gesetze-im-internet.de*) as a three-level hierarchy: *Gesetzbuch* (statute) → *Abschnitt* (section) → *Paragraf* (provision). Intra-law cross-references are detected by regex and materialised as *referenziert* edges.

The KGLawLinker connects the domain KG to the law graph via three tiers of decreasing confidence: (1) *Explicit citation parsing* (confidence ≥ 0.95): regex extraction of statutory references (e.g., § 7 Abs. 3 AtG) from entity evidence text; (2) *Keyword-to-provision mapping* (0.70–0.85): ~ 35 domain-specific keyword patterns mapped to known regulatory provisions; (3) *Entity-type defaults* (0.45–0.60): class-level fallback links when no explicit citation or keyword match exists. On the nuclear corpus, the resulting law graph contains 858 nodes and 3,684 edges with 3,078 intra-law cross-references.

4. Evaluation

4.1. Setup

Models and infrastructure All experiments use locally-served LLMs via Ollama 0.18. Primary model: gemma4:e2b (Google Gemma 4, 2B parameters) [12]. Additional conditions use gemma4:e4b (4B), gemma4:31b (31B), and nemotron-3-nano (4B) to measure scale and architecture effects. Embeddings: qwen3-embedding (4096-dim). Full stack on one workstation via Docker Compose; no external network access during experiments.

Dataset *Nuclear decommissioning corpus*: 33 documents (safety reports, licence applications, decommissioning descriptions), 3,004 text segments, ontology with 58 OWL classes and 120 object properties. *German law corpus*: 17 federal statutes parsed from structured XML into 5,857 paragraph nodes—used as a domain-transfer demonstration and as law-graph input for Condition C.

Conditions Six conditions ($n = 3$ runs each): A (single-pass, 2B), B (5-pass, 2B), C (5-pass + law graph, 2B), D (5-pass, 4B), E (5-pass, 31B), F (5-pass, nemotron-nano). Key comparisons: **A vs. B** (iterative discovery), **B vs. C** (law graph), **B vs. D/E/F** (model scale/architecture).

4.2. Results

Table 1

Graph quality metrics (mean $\pm$ std, $n = 3$; F: $n = 2$). $\uparrow$ higher is better; best per row in **bold**.

	A	B	C	D	E	F
Class cov. (%) $\uparrow$	33.7 $\pm$ 4.0	35.7 $\pm$ 3.8	39.7$\pm$1.5	24.3 $\pm$ 0.6	22.0 $\pm$ 0.0	36.5 $\pm$ 0.7
Rel. cov. (%) $\uparrow$	41.3 $\pm$ 9.8	74.3 $\pm$ 9.0	104.5$\pm$8.1	54.0 $\pm$ 3.3	44.5 $\pm$ 1.8	30.5 $\pm$ 4.2
Entities $\uparrow$	377 $\pm$ 12	572 $\pm$ 42	612$\pm$22	568 $\pm$ 14	461 $\pm$ 8	240 $\pm$ 4
Relations $\uparrow$	328 $\pm$ 54	613 $\pm$ 37	617 $\pm$ 51	821$\pm$23	680 $\pm$ 26	264 $\pm$ 16
Avg. degree $\uparrow$	3.3 $\pm$ 1.7	7.2 $\pm$ 1.7	12.6 $\pm$ 1.6	10.2 $\pm$ 2.8	13.5$\pm$3.0	10.4 $\pm$ 1.3
Orphan rate (%) $\downarrow$	21.9 $\pm$ 16.5	12.6 $\pm$ 1.7	8.9$\pm$2.5	17.1 $\pm$ 2.1	13.0 $\pm$ 2.2	16.5 $\pm$ 1.7
SHACL score $\uparrow$	0.937 $\pm$ 0.011	0.942 $\pm$ 0.010	0.953$\pm$0.004	0.913 $\pm$ 0.002	0.907 $\pm$ 0.000	0.944 $\pm$ 0.002
Build time (min)	37.6$\pm$2.2	75.8 $\pm$ 3.4	58.4 $\pm$ 4.6	76.6 $\pm$ 1.4	218 $\pm$ 5	2732 $\pm$ 457

Key findings

- Iterative discovery (A $\rightarrow$ B) substantially improves graph richness:** entity count +51.7% (377 $\rightarrow$ 572), relation coverage +33.0 pp, orphan rate -9.4 pp. Class coverage rises only modestly (+2.0 pp), confirming that iteration primarily deepens relation population rather than discovering entirely new classes.
- Law-graph augmentation (B $\rightarrow$ C) strengthens connectivity and conformance:** average degree 7.19 $\rightarrow$ 12.58, orphan rate 12.6 $\rightarrow$ 8.9%, SHACL score 0.942 $\rightarrow$ 0.953. Legal context primarily adds cross-entity relations that raise structural coherence.
- Larger models do not improve ontology-aligned quality:** 31B (E) trails the 2B multi-pass baseline on class coverage (22.0 vs. 35.7), relation coverage (44.5 vs. 74.3), and SHACL score (0.907 vs. 0.942) at 2.9 $\times$ the build time. The 4B model (D) extracts more raw relations but achieves lower ontology-aligned metrics. For cost-effective, ontology-consistent KG extraction, multi-pass 2B (B/C) offers the best quality–cost trade-off.
- All conditions achieve zero SHACL hard violations**, confirming that constraint-guided extraction enforces validity by construction.

Limitations The current evaluation measures structural graph quality; factual correctness of extracted triples is not yet assessed against a manually annotated gold set. Future work will add LLM-as-a-judge scoring and manual annotation to complement structural metrics. The corpus (33 documents) is sufficient for ablation but limits broad generalisation claims.

5. Conclusion

We presented an iterative, ontology-driven pipeline for KG construction from unstructured documents, operating entirely on local LLMs. The CQ-driven discovery loop, six-stage confidence pipeline, and integrated SHACL validation together deliver consistent improvements over single-pass baselines—without requiring larger models or cloud APIs. Domain transferability is demonstrated on two fundamentally different domains using identical pipeline code.

Future directions include factual correctness evaluation against annotated benchmarks, co-evolutionary ontology–KG construction where coverage gaps trigger schema refinements, and scaling to larger corpora via parallel GPU extraction.

Declaration on Generative AI

During the preparation of this work, the authors used GitHub Copilot for coding assistance (code completion, refactoring, debugging) and ChatGPT/Claude for grammar and style corrections. All AI-assisted content was reviewed and edited by the authors, who take full responsibility for the publication’s content.

References

- [1] Hogan A *et al.* Knowledge Graphs. *ACM Computing Surveys*. 2021;54(4):1–71.
- [2] Ji S *et al.* A Survey on Knowledge Graphs: Representation, Acquisition, and Applications. *IEEE Trans. Neural Netw. Learn. Syst.* 2022;33(2):494–514.
- [3] Bian H *et al.* LLM-Empowered Knowledge Graph Construction: A Survey. arXiv:2510.20345. 2025.
- [4] Pan S *et al.* Unifying Large Language Models and Knowledge Graphs: A Roadmap. *IEEE Trans. Knowl. Data Eng.* 2024;36(7):3580–3599.
- [5] Etzioni O *et al.* Open Information Extraction from the Web. *Commun. ACM*. 2008;51(12):68–74.
- [6] Feng X, Miao S. Ontology-Grounded Automatic Knowledge Graph Construction under Wikidata Schema. arXiv:2412.20942. 2024.
- [7] Zhu Y *et al.* AutoKG: Knowledge Graph Construction with LLMs and Multi-Agent Systems. arXiv:2305.13168. 2024.
- [8] Grüninger M, Fox MS. Methodology for the Design and Evaluation of Ontologies. *Proc. Workshop Basic Ontological Issues in Knowledge Sharing (IJCAI-95)*. 1995.
- [9] Shimizu C, Hitzler P. Accelerating Knowledge Graph and Ontology Engineering with LLMs. arXiv:2411.09601. 2024.
- [10] Li W, Poveda-Villalón M, Garijo D. Large Language Models for Ontology Engineering: A Systematic Literature Review. arXiv:2404.03044. 2024.
- [11] Robertson S, Zaragoza H. The Probabilistic Relevance Framework: BM25 and Beyond. *Found. Trends Inf. Retr.* 2009;3(4):333–389.
- [12] Gemma Team. Gemma: Open Models Based on Gemini Research and Technology. arXiv:2403.08295. 2024.
- [13] Bundesministerium der Justiz. Gesetze im Internet. <https://www.gesetze-im-internet.de/>. 2026.