AIDOC-AP: An Application Profile for Technical Documentation of AI Systems

Journal Title
XX(X):1–13
©The Author(s) 2026
Reprints and permission:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Sebastian Neumaier[1,2], Tobias Dam[1], and Fabian Kovac[1]

Abstract

We present AIDOC-AP, an application profile for representing technical documentation of AI systems in accordance with technical documentation requirements of the EU AI Act. Our methodology adapts the NeOn-GPT pipeline, an ontology engineering framework powered by large language models (LLMs), and extends it with two steps: (i) LLM-assisted alignment to reference ontologies and (ii) LLM-based iterative coverage evaluation of competency questions. The resource comprises an OWL ontology for AI system lifecycle documentation and an ontology of Annex IV requirements and competency questions. We validate AIDOC-AP on three pilot use cases from the CERTAIN project spanning distinct AI Act high-risk domains, including an empirical evaluation of competency-question answering over the instantiated knowledge graphs, and publish all artifacts under open licenses.

License: CC BY 4.0
DOI: 10.5281/zenodo.17787786
URI: https://w3id.org/aidoc-ap

Keywords

EU AI Act, Ontology, Knowledge Graph, Technical Documentation, Compliance, AI System

Introduction

Recent regulatory developments such as the EU AI Act[11*] impose detailed technical documentation obligations on providers of high-risk AI systems. The primary obligation is established by Article 11(1) of the AI Act, which requires providers to draw up technical documentation before a high-risk AI system is placed on the market and to keep it up to date; Annex IV specifies the minimum elements this documentation shall contain, including comprehensive information on system architecture, data, training and testing procedures, risk management and lifecycle performance. Notably, the Commission is empowered to amend Annex IV by delegated acts (Article 11(3)), so any supporting resource must be designed to adapt to an evolving requirements list. However, current documentation practices are mostly unstructured and text-based, making automated compliance checking and cross-system comparison difficult.

This paper introduces AIDOC-AP, an ontology-based application profile to represent AI system documentation in a structured, machine-readable way. AIDOC-AP is explicitly driven by AI Act documentation requirements (in particular, the AI Act Annex IV requirements) and builds on established vocabularies. Methodologically, we adapt the NeOn-GPT pipeline[13], a recent LLM-powered ontology engineering framework, and extend it with two steps that use large language models for (i) assisting alignment to reference ontologies and (ii) iterative coverage evaluation of competency questions (CQs).

From a resource perspective, we contribute the following artifacts:

1. **AIDOC-AP**, an application profile for Annex IV technical documentation, implemented as an OWL ontology.
2. **Annex IV Requirements**, formalising the requirements of Annex IV and 50 competency questions, each traceably linked to the underlying legal text.
3. **Semantic alignments** between AIDOC-AP and established reference ontologies (e.g., AIRO, VAIR, DPV, MLSchema, MCRO), including structural similarity scores and curated SKOS mapping relations (`skos:exactMatch`, `closeMatch`, `broadMatch`, `narrowMatch`), complemented by `rdfs:subClassOf` specialisations in the ontology itself.
4. **An extended ontology engineering pipeline**, adding two novel LLM-assisted steps, i) semantic alignment and ii) CQ coverage evaluation, to the NeOn-GPT

[1]University of Applied Sciences St. Pölten, AT
[2]Vienna University of Economics and Business, AT

Corresponding author:
Sebastian Neumaier, University of Applied Sciences St. Pölten, Data Intelligence Research Group, St. Pölten, Austria
Email: sebastian.neumaier@ustp.at

*All article, paragraph and annex references in this paper refer to Regulation (EU) 2024/1689 as originally published[11]. The Regulation has since been amended by Regulation (EU) 2026/1744[12], which inserts new articles and modifies several existing provisions (including Article 11(1), which now allows SMEs and small mid-caps to provide the Annex IV elements in simplified form); references should therefore be checked against the consolidated text. The list of Annex IV documentation elements modelled in this work is not changed by the amendment.

methodology, implemented in an openly available workflow.

5. **A cross-domain validation** instantiating application-profile-compliant knowledge graphs for three AI systems in distinct AI Act high-risk domains, with an empirical, SPARQL-based evaluation of Annex IV competency-question answering.

The remainder of this paper is structured as follows: Section Background and Related Work discusses background and related work. Section Methodology presents our extended NeOn-GPT methodology. Section Annex IV Requirements and Competency Questions formalises Annex IV requirements as ontology concepts and competency questions. Section AIDOC-AP Ontology presents the resulting AIDOC-AP ontology. Section Results: Alignments and Coverage provides ontology alignments with reference vocabularies and presents the evolution of Annex IV coverage across ontology iterations. Section Use Case Validation reports on validation across three real-world pilot use cases, Section Discussion discusses limitation of our approach, and Section Conclusion and Future Work concludes.

Background and Related Work

Technical documentation for AI systems is largely shaped by textual artefacts such as model cards[26] and datasheets for datasets[14]. These formats provide rich human-readable descriptions but typically lack the machine-readable structure required for automated assessments of completeness or compliance with regulations such as the EU AI Act.

Methodologically, our work is motivated by the gap between such narrative documentation and regulation-aware representations. We seek a representation that (i) is explicitly grounded in Annex IV, (ii) supports automated reasoning and validation, and (iii) remains compatible with existing documentation practices.

Ontologies for AI Systems, Risk and Data

We review ontologies that provide concepts for AI systems, risk management, datasets and data quality, including AIRO[17], VAIR[18], MLSchema[10], MCRO[4], RAINS[28], MLSO[8], DCAT[1], DQV[2], PROV-O[22]. Methodologically, we position AIDOC-AP as an application profile that reuses these vocabularies where possible and introduces new terms based on the Annex IV structure.

We structure the comparison along four dimensions: (i) support for system architecture and components, (ii) coverage of data requirements and dataset characteristics, (iii) support for AI lifecycle stages, and (iv) explicit grounding in the Annex IV technical documentation requirements. A dimension is marked as covered (✓) if the ontology provides dedicated classes or properties for it, as partially covered ((✓)) if it provides related concepts without addressing the dimension's full scope, and as not covered (✗) otherwise. Note that dimension (iv) is deliberately narrower than regulatory grounding in general: AIRO and VAIR are explicitly grounded in the AI Act—AIRO was, to our knowledge, the first ontology to model AI Act concepts, with definitions referencing specific clauses—but neither specifically targets the structure of the Annex IV documentation items.

Table 1. Comparison of existing ontologies along four dimensions. (✓) denotes partial coverage.

Ontology	System Architecture	Data Req. and Dataset	Lifecycle Coverage	Annex IV Grounding
MLSchema	✓	✓	✓	✗
MCRO	✓	✓	✓	✗
AIRO	✗	(✓)	✓	✗*
VAIR	✓	(✓)	✓	✗*
DPV	✓	(✓)	✓	(✓)*
MLDCAT-AP	(✓)	✓	(✓)	✗
DCAT3	✗	✓	✗	✗
PROV-O	✗	✗	✓	✗
AIDOC-AP	✓	✓	✓	✓

*AIRO and VAIR are grounded in the AI Act in general but do not model Annex IV documentation items. DPV's AI Act extension *does* provide dedicated Annex IV documentation concepts (e.g. `TechnicalDocumentation`, `SystemArchitectureDescription`, `TestLog`, `EUDeclarationOfConformity`); AIDOC-AP differs by decomposing Annex IV into requirement-linked competency questions with coverage tooling, rather than atomic documentation concepts.

Table 1 summarises how existing ontologies differ across these dimensions and highlights where AIDOC-AP provides the missing combinations. We exclude MEX, RAINS and ML-Onto from the table, to avoid redundancy: MEX overlaps with PROV-O's broader provenance coverage, RAINS overlaps with the risk management concepts represented by AIRO and VAIR, and MLSO focuses on academic ML task classification rather than regulatory documentation.

In contrast to AIRO and VAIR, which model AI system concepts and risk structures, AIDOC-AP builds on this groundwork by anchoring each term to Annex IV and by representing the documentation artefacts required for compliance. In addition, AIDOC-AP provides LLM-assisted tooling for semantic alignment and Annex IV coverage assessment, enabling practical compliance-oriented KGs rather than purely conceptual representations.

The term *application profile* and the "-AP" naming follow the convention established by the SEMIC community, most prominently DCAT-AP[29], which has recently been extended to MLDCAT-AP[21] for describing machine learning models and datasets as catalogue entries. AIDOC-AP is complementary to MLDCAT-AP: while MLDCAT-AP targets the discoverability of AI/ML assets in data catalogues, AIDOC-AP targets the structured representation of the Annex IV lifecycle documentation itself; a system may reasonably employ both. Beyond catalogue-oriented profiles, the Data Privacy Vocabulary (DPV)[31] provides an actively maintained family of vocabularies for EU regulations, including a TECH extension for technologies, an AI extension, and a dedicated AI Act extension.[†] DPV covers many of the concepts appearing in AI Act compliance scenarios

[†] https://w3id.org/dpv/2.3/tech/, https://w3id.org/dpv/2.3/ai/, https://w3id.org/dpv/2.3/legal/eu/aiact/

across multiple regulations, and its AI Act extension already includes dedicated Annex IV documentation concepts (e.g. `TechnicalDocumentation`, `SystemArchitectureDescription`, `TestLog`, `EUDeclarationOfConformity`). AIDOC-AP does not duplicate these but complements them with structural depth—decomposing Annex IV into requirement-linked competency questions with coverage tooling—and grounds its Annex IV-specific terms in DPV concepts where available (Section Results: Alignments and Coverage) to maximise interoperability with other DPV-based compliance approaches. Closest to our goal, Golpayegani et al.[19] propose semantic frameworks to support implementation of the EU AI Act, including its documentation and registration obligations, and template-based approaches to AI Act technical documentation have been proposed as well[25]. AIDOC-AP differs in its Annex IV structural depth: rather than templates or high-level frameworks, it provides requirement-linked competency questions, lifecycle stage mappings, and tooling for iterative and empirical coverage evaluation of the documentation itself.

Recent work in RegTech and legal informatics has explored modelling regulatory obligations and compliance processes as knowledge graphs[6,16,20]. These approaches aim to formalise norms, obligations and evidence structures, including emerging efforts to represent the EU AI Act. However, they typically focus on high-level regulatory logic rather than the concrete content of technical documentation.

LLM-assisted Ontology Engineering

Recent work integrates LLMs into ontology engineering workflows. NeOn-GPT[13] automates term extraction, definition generation, taxonomy building, and axiom suggestion, while combining reasoning and manual review. Other approaches explore LLM-based ontology matching: LLMs4OM uses GPT for zero-shot alignment across ontologies[5], and more complex alignment tasks have been tackled by prompt-based frameworks[3] or multi-agent dialogue with LLMs[36]. There is also work using LLMs to generate competency questions for validating ontology design, such as identifying semantic modelling pitfalls[7].

Complementary lines of work address the automated verification of competency questions against ontologies—for example by translating CQs into SPARQL-OWL test queries[35]—and the (increasingly LLM-assisted) evaluation of ontologies against such requirements[23]. Our coverage step builds on this idea but applies it in a regulation-driven loop, and we additionally validate the resulting ontology by executing the competency questions as SPARQL over instantiated pilot graphs (Section Use Case Validation). Our contribution is not any individual technique but their integration: an ontology engineering pipeline in which LLM-assisted alignment to reference ontologies and LLM-based, CQ-driven coverage estimation are combined as recurring steps of a regulation-driven development loop, with LLM outputs subject to automated reasoning and human curation (Section Methodology). To our knowledge, no prior pipeline integrates both steps in this iterative, requirements-tracking manner.

Methodology

This section describes the ontology engineering pipeline used to develop AIDOC-AP. We first motivate the choice of NeOn-GPT as the base methodology and give an overview of the extended pipeline, then summarise the five NeOn-GPT steps, and finally detail the two extensions this work adds: LLM-assisted alignment to reference ontologies (step 6) and LLM-based iterative coverage evaluation against the Annex IV requirements (step 7).

Overview of the Extended NeOn-GPT Pipeline

Our ontology engineering methodology is based on the NeOn-GPT pipeline[13], which integrates large language models into competency-question-driven ontology development. The base pipeline comprises five steps: (1) extraction of domain-specific terminology from source documents, (2) LLM-assisted generation of term definitions and taxonomic relations, (3) LLM-assisted axiomatisation (domain/range constraints, cardinality restrictions, disjointness), (4) automated reasoning and (5) pitfall detection.

We chose NeOn-GPT over alternative ontology engineering approaches for three reasons: it is competency-question-driven, which matches our requirement-centric setting where every term must trace back to an Annex IV item; it follows the established NeOn methodology and therefore yields a documented, repeatable process rather than ad-hoc LLM usage; and its step structure made it straightforward to insert the two additional steps we require (alignment and coverage evaluation). Among the alternatives we considered, classical frameworks such as LOT[32] provide a mature engineering process but no LLM assistance, dataset-driven generators such as OntoGenix[34] start from data rather than from regulatory requirements, and matching-only approaches in the style of LLMs4OM[5] cover only the alignment task rather than the full development loop. We use NeOn-GPT as a structured prompting pipeline with defined per-step prompts and outputs; its conversational/interactive aspect was not used.

We extend this pipeline with two additional steps that address the challenges of building a regulation-driven application profile: (6) LLM-assisted alignment to reference ontologies, and (7) LLM-based iterative coverage evaluation of competency questions. These extensions help us to integrate existing vocabularies and to track how well the ontology supports the requirements it is intended to cover. Figure 1 details these steps of our pipeline; steps 1-5 are based on the NeOn-GPT methodology.

Steps 1–5: Based on NeOn-GPT Pipeline

(1) Terminology Extraction and Definition Generation. We begin by extracting candidate terms from Annex IV and from existing AI documentation frameworks (AIRO ontology, model cards, datasheets). For each term, we prompt an LLM to generate a short definition and to suggest synonyms or related concepts. We, the authors of the ontology, review and refine these definitions to ensure alignment with the initial legal text and domain practices.

(2) Taxonomy Construction. Using the refined term list, we prompt the LLM to propose subclass relations. For

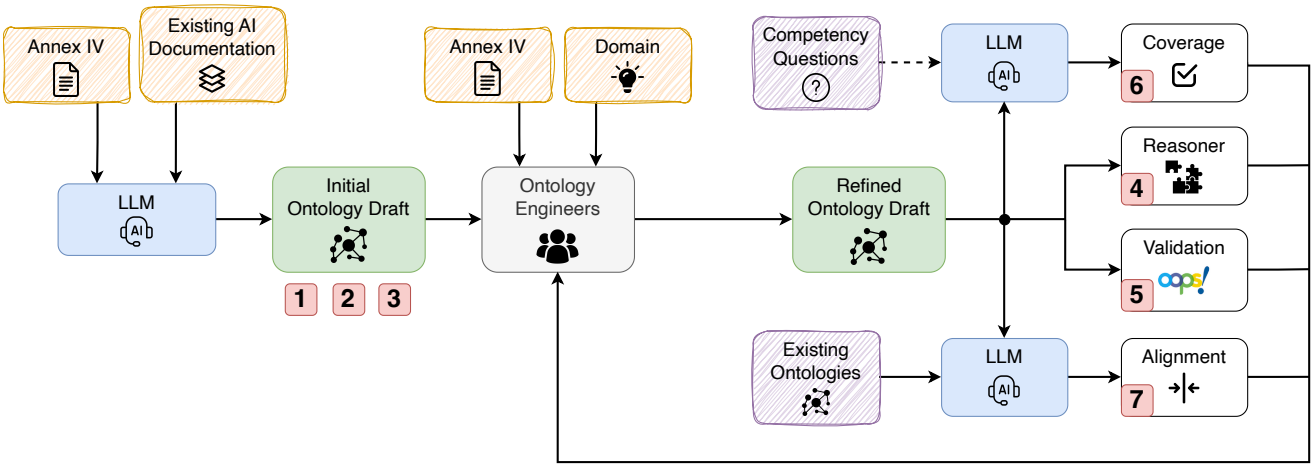


Figure 1. Extended NeOn-GPT pipeline for AIDOC-AP development. The numbers refer to the methodological step described in Section Methodology.

example, `aidoc:TrainingDataSheet` is suggested as a subclass of `aidoc:DataSheet`. We again validate the proposed subsumptions and adjust the taxonomy so that it supports the required queries.

(3) Axiomatisation. We then prompt the LLM to suggest OWL axioms, including domain and range restrictions. E.g., the property `aidoc:usesTrainingData` is constrained to have domain `aidoc:DataTraining` and range `aidoc:Dataset`. All suggested axioms are reviewed by ontology engineers before being integrated into the OWL file.

(4) Automated Reasoning & (5) Pitfall Detection. After each modelling iteration, we run the HermiT reasoner to check for logical inconsistencies (e.g., unsatisfiable classes, conflicting axioms). Any detected inconsistencies are resolved by adjusting axioms or class hierarchies.

We periodically run the OOPS! validator to identify common ontology pitfalls, such as incomplete annotations or ambiguous naming. Detected pitfalls are reviewed and, where appropriate, corrected.

Step 6: LLM-assisted Alignment to Reference Ontologies

To ensure interoperability with existing Semantic Web vocabularies, we systematically align AIDOC-AP to reference ontologies such as PROV-O, MLSchema, DCAT, AIRO, VAIR and DPV (including its AI and AI Act extensions). Manual alignment is time-consuming and error-prone, particularly when the reference ontologies are large; we therefore introduce an LLM-assisted alignment step.

Alignment methodology. The alignment implements a classification function

$$g : (t_{\text{aidoc}}, t_{\text{ref}}) \mapsto (r, p),$$

where $(t_{\text{aidoc}}, t_{\text{ref}})$ is a pair of ontology terms (with their labels and descriptions), r is a SKOS mapping relation drawn from a small, fixed set (`skos:exactMatch`, `skos:closeMatch`, `skos:broadMatch`, `skos:narrowMatch`, `n/a`), and $p \in [0, 1]$ is a confidence score.

Implementation. We first generate candidate pairs using lexical similarity over labels and comments (based on normalised Levenshtein distance and token overlap). Since reference files redeclare terms from other vocabularies, the same pair can surface under several files; such duplicates are merged and each unique pair is assigned to the vocabulary owning the target IRI. For each candidate pair, we construct a prompt that presents the LLM with the labels, definitions and usage contexts of both terms, together with explicit definitions of the admissible SKOS mapping relations, and ask it to classify the relationship and assign a score p. We emphasise that p is not treated as a calibrated probability—LLMs are known to produce unreliable self-assessed confidence values—but solely as a heuristic to rank and pre-filter candidates for human review: only correspondences with p above a fixed threshold (0.75) are retained in the semantic alignment files, and *every* retained triple is reviewed by domain experts, who accept, reject or refine the suggested relation. We report the outcome of this curation quantitatively in Section Results: Alignments and Coverage, including the precision of LLM suggestions and an analysis of a curated sample of pairs below the lexical candidate-generation threshold to estimate false negatives.

Example alignment prompt (simplified): ‡

> You are an experienced knowledge engineer.
> **AIDOC-AP term:** `aidoc:Dataset`
> Label: “Dataset”
> Definition: A collection of data used for training, validation or testing.
> **Reference term:** `dcat:Dataset`
> Label: “Dataset”
> Definition: A collection of data, published or curated by a single agent.
> **Task:** Classify the relationship between these two terms. Choose one:
> `skos:exactMatch`, `skos:closeMatch`, `skos:broadMatch`, `skos:narrowMatch`, or `n/a`

‡The full prompt is available in the project repository.

(no meaningful relation). Assign a confidence score between 0.0 and 1.0.
Return a JSON object with fields: relation, confidence, explanation.

Step 7: LLM-based Iterative Coverage Evaluation

A key challenge in building a regulation-driven ontology is ensuring that it adequately covers all legal requirements. We operationalise this notion of "coverage" through competency questions: for each Annex IV requirement, we define a set of CQs that a knowledge graph must be able to answer. We then use an LLM-based evaluator to estimate how well the current ontology supports these CQs.

Coverage evaluation as an iterative process. The evaluator is run after each significant ontology revision (e.g., after adding a new module or refining axioms following pitfall detection). Coverage scores are compared across iterations, allowing us to track progress and identify requirements that remain under-specified. This iterative feedback loop guides ontology refinement: low-scoring requirements prompt us to add missing classes, properties or axioms.

Formal definition. The evaluator implements a function

$$f : (r, O) \mapsto (S_r, c_r, e_r)$$

where r is a textual requirement description, O is the current ontology, S_r is a set of suggested AIDOC terms, $c_r \in [0, 1]$ is a scalar coverage score, and e_r is a short natural-language explanation.

Input and output schema. As input, the evaluator receives (i) the plain-text Annex IV requirement and associated competency questions, and (ii) a catalogue of candidate AIDOC entities with labels and comments. As output, it returns a JSON object per requirement with fields: `"id"` (requirement identifier), `"score"` (c_r), `"matched_terms"` (S_r as labels), and `"explanation"` (e_r). We set the temperature to 0 for all models to minimise sampling stochasticity, and repeat each evaluation with three different seeds to quantify the remaining run-to-run variance (reported in Section Results: Alignments and Coverage). Each requirement is evaluated in an independent, single-turn request: no conversation state is carried over between requirements, between ontology iterations, or between runs, so successive iterations cannot influence each other through accumulated context. The models used at each pipeline stage, including references to their model cards, are summarised in Table 2.

Metric interpretation. The coverage score c_r is an *LLM-estimated indicator*, not a validated metric or ground truth assessment. It is defined as a subjective estimate, provided by the LLM under explicit instructions, of how well the current ontology supports answering the competency questions associated with requirement r. We constrain c_r to the unit interval and ask the model to justify high and low scores by referring to concrete classes and properties. To assess whether these estimates are meaningful, we validate them in two ways: (i) we repeat the coverage computation across three independent seeded runs per model and report the run-to-run variance, which stays below 0.009 (Section Results: Alignments and Coverage), and (ii) we compare the LLM-estimated scores against an independent expert assessment: for a seeded random sample of ten requirements, one author scored the iteration-3 ontology against the same term catalogue and score definition, blind to the LLM outputs. The pooled LLM scores deviate from the expert scores by 0.12 on average (mean absolute deviation), and seven of the ten requirements fall into the same low/medium/high coverage band; notably, the three larger deviations are all cases in which the LLM scores *higher* than the expert, consistent with the upward bias discussed in Section Discussion (sample sheet and analysis script in the repository). However, this score should be interpreted as a decision-support estimate for guiding ontology development, not as a compliance guarantee.

Example coverage evaluator prompt (simplified): [§]

You are an experienced knowledge engineer and expert on the EU AI Act.
Requirement: "A detailed description of the system architecture, including algorithms. . . "
Competency questions: "What is the architecture and what algorithms are used? Which software components make up the system?"
Ontology terms (label – description):
`aidoc:AISystem` – An artificial intelligence system.
`aidoc:SoftwareComponent` – A software module or service.
`aidoc:hasArchitecture` – Textual description of system architecture.
. . .
Task: Select at most 10 ontology terms that best capture the information needed to satisfy this requirement and its questions. Then assign an overall coverage score between 0.0 (no coverage) and 1.0 (excellent coverage). Explain your choice in 3–5 sentences.
Return a JSON object with fields: `id`, `score`, `matched_terms`, `explanation`.

Annex IV Requirements and Competency Questions

We formalise Annex IV as an ontology, where each requirement is represented as an instance of `aiact:Requirement` within a SKOS concept scheme. The ontology encodes requirement identifiers, labels, textual descriptions and lifecycle annotations.

We based our modelling on the text of the EU AI Act as published in 2024, using the official Annex IV [11]. Each numbered item in Annex IV was mapped to a distinct requirement individual (`aiact:req1`–`aiact:req22`). Where a textual item contained multiple logically separable obligations, we captured them as separate competency questions attached to a single requirement (cf. Section Competency Questions). For each requirement, we preserved a trace to the original legal text through identifiers and human-readable labels.

[§] The full prompt is available in the project repository.

Table 2. Open-weight LLMs used in the pipeline. All models are served locally via Ollama at temperature 0 with fixed seeds (43–45), which makes the full pipeline reproducible without proprietary APIs. Quantisation is the format actually served on our infrastructure.

Model	Quantisation	Pipeline stage	Rationale
Gemma 3 27B[15]	Q4_K_M	alignment + coverage	strong instruction following at a size that runs locally
Llama 3.3 70B[24]	Q4_K_M	coverage	widely used open baseline
GPT-OSS 120B[30]	MXFP4	coverage	large open reasoning model, different model family
Apertus 70B[33]	Q4_K_M	coverage	fully open (weights and training data), European model

The resulting ontology module, published online[¶] and as `TTL` file, currently comprises 22 requirement individuals and 50 competency questions, all organised in a single SKOS concept scheme `aiact:`. Besides labels and descriptions, each requirement is annotated with an `aiact:aiLifecycleStage` string that links it to data, machine-learning or software pipeline phases, as well as cross-cutting activities.

Competency Questions

For each requirement we define one or more competency questions (CQs) that describe the information a knowledge graph must provide to demonstrate coverage. We represent CQs as instances of `aiact:CompetencyQuestion` linked to requirements via `aiact:hasCompetencyQuestion`. The derivation process for CQs is described in Section Methodology.

The CQ set is intentionally biased towards questions that can be operationalised on technical documentation artefacts (architecture diagrams, data sheets, logs) rather than purely legal concepts. This improves the practical usefulness of AIDOC-AP for system documentation, but means that some normative or organisational aspects of Annex IV remain outside the scope of the current CQs. Moreover, the paraphrasing from legal text to CQs introduces a degree of subjectivity that may affect subsequent coverage evaluations; to mitigate this, we keep CQs close to the wording of the original requirements and maintain explicit links between each CQ, its parent requirement and the ontology modules it touches. Future work will include expert review of the CQ set by legal and compliance professionals and a comparison with alternative CQ formulations.

Table 3 shows how selected Annex IV requirements map to competency questions and AIDOC-AP modules. For example, *req11* (Data requirements) is operationalised through five CQs that query classes such as `DataSheet`, `Labeling Procedure`, and `DataCleaningProcedure`, enabling systematic verification of documentation completeness.

AIDOC-AP Ontology

Typically, an *application profile (AP)* is a specification that constrains and combines terms from multiple existing vocabularies for a particular use case, adding a minimal set of new terms only where required, and prescribing how these terms should be used together through explicit constraints, alignments, and machine-readable implementation guidance.

AIDOC-AP follows the DCAT-AP tradition[29]: it prioritises interoperability through vocabulary reuse, lightweight extensions, clear modelling constraints, and formal alignments to established standards.

As of now, AIDOC-AP is an EU AI Act Annex IV–focused application profile: rather than defining a standalone ontology, AIDOC-AP introduces locally named terms in its own namespace and grounds them in semantically appropriate terms from established vocabularies—as subclasses where it specialises them, and via SKOS mappings otherwise. This allows the profile to adopt consistent naming and enforce Annex IV–specific constraints while reusing the semantics of external ontologies.

Reused Vocabularies. AIDOC-AP reuses the semantics of established vocabularies through subclass and SKOS-mapping axioms. The following vocabularies are integrated:

- **PROV-O** for provenance of data, models, activities, and agents;
- **DCAT** for dataset and distribution descriptions;
- **MLSchema** for models, training processes, and ML experiments;
- **AIRO** for risk management and assessment artefacts;
- **VAIR** for AI system characterisation and documentation elements;
- **DQV** for performance metric definitions (minimal usage: `PerformanceMetric` as subclass of `dqv:Metric`);
- **DPV** (core, AI, AI Act and TECH extensions) for grounding AIDOC-AP terms in the actively maintained EU-regulation vocabularies; e.g. `aidoc:AISystem` is declared `rdfs:subClassOf dpv-ai:AISystem`, and the provider/deployer roles subclasses of the corresponding DPV AI Act concepts.

Profile Constraints and External Grounding. AIDOC-AP grounds its locally named terms in established vocabularies at two levels:

- **Logical specialisation.** When an AIDOC-AP term narrows an external term to the Annex IV documentation context—e.g. by mandatory documentation artefacts or cardinality constraints—it is declared `rdfs:subClassOf` (respectively `rdfs:subPropertyOf`) of that term. Specialisation is the semantically correct relation: adding constraints *narrows* the extension, so the AIDOC-AP class is a subclass of, not co-extensional with, the external class. For

[¶] https://certain-project.github.io/aidoc-ap/requirements.html

Table 3. Mapping of Requirements to Competency Questions and example AIDOC-AP Modules.

Requirement	Competency Questions	AIDOC-AP Concepts
General description (Annex IV, Section 1(a))	**CQ1.1:** What is the intended purpose? **CQ1.2:** What is the provider name? **CQ1.3:** What is the system version?	`AISystem`, `AIProvider`, `IntendedPurpose`, `version`
System architecture (Annex IV, Section 2(c))	**CQ10.1:** What architecture and algorithms are used? **CQ10.2:** Which components make up the system? **CQ10.3:** How do components integrate? **CQ10.4:** Which computational resources are used?	`SoftwareComponent`, `hasComponent`, `feedsIntoComponent`, `ComputationalResource`
Validation & testing (Annex IV, Section 2(g))	**CQ14.1:** What validation procedures are used? **CQ14.2:** What validation/testing data is used? **CQ14.3:** What metrics measure accuracy/robustness? **CQ14.4:** Are test logs and reports available?	`ModelEvaluation`, `DataValidation`, `PerformanceMetric`, `usesValidationData`, `Log`

example, `aidoc:AISystem` is a subclass of `dpv-ai:AISystem`.

- **Interoperability mappings.** Correspondences that do not follow a clean subsumption (concepts that merely coincide, or where the external term is the narrower one) are published as SKOS mapping relations (`skos:exactMatch`, `closeMatch`, `broadMatch`, `narrowMatch`) in separate alignment files (Section Results: Alignments and Coverage), linking concepts without committing to class-level logical entailment. We use the mapping properties because these links cross vocabularies, following the SKOS convention for relating concepts across concept schemes (`skos:broadMatch` being a sub-property of `skos:broader`); hierarchical links within AIDOC-AP's own vocabularies use `skos:broader`.

The two levels are connected through curation: the alignment step (Section Methodology) proposes concept-level SKOS mappings, and where a curated correspondence reflects a genuine specialisation—AIDOC-AP being the narrower, constrained term—the ontology asserts the corresponding `rdfs:subClassOf`. This is also where an over-strong LLM proposal is refined: for `aidoc:AISystem` the model suggested an "exact" match, whereas a curator, aware that the Annex IV constraints make it a subclass rather than an equivalent of `dpv-ai:AISystem`, corrects the grounding accordingly—a concrete illustration of why human curation is retained.

We deliberately avoid `owl:equivalentClass`: it asserts extensional co-equality[9], which the Annex IV-specific constraints AIDOC-AP adds would violate. Keeping logical subsumption in the ontology and looser correspondences as SKOS mappings also keeps reasoning monotonic and side-effect-free, since AIDOC-AP does not import the external ontologies wholesale.

Classes vs. string-valued properties. Provider documentation typically records many items as free text, and the most direct translation into an ontology would be string-valued properties. This, however, forfeits semantic value: two string values cannot be related, compared, or organised hierarchically. We therefore apply a simple modelling criterion: where a piece of information denotes a concept that can be specialised or instantiated, or that needs to be related to other concepts, it is modelled as a class; where it is purely descriptive, it is kept as a string literal. Following this criterion, the following concepts are modelled as classes with a controlled vocabulary: the intended purpose (`aidoc:IntendedPurpose`, aligned with the corresponding DPV AI Act concept) and the system's form of provision (`aidoc:Modality`: standalone software, software service/API, embedded component, hardware product, cloud service), each linked through an object property (`aidoc:hasIntendedPurpose`, `aidoc:hasModality`). This lets values be linked (e.g., via `skos:broader`/`narrower`), compared and tracked across system versions—which matters under the AI Act, where a changed intended purpose can alter obligations. We deliberately provide no parallel string-valued shortcuts for these concepts, leaving a single, unambiguous way to express each of them. Purely descriptive information (e.g., free-text labelling guidelines, human-readable descriptions) is kept as strings; version identifiers are currently strings as well, although they could be promoted to classes should version-level relations become necessary.

Scope and Non-Scope Decisions. AIDOC-AP is intentionally scoped to artefacts required under the AI Act documentation requirements. The profile:

- *includes* only entities, processes, and artefacts directly connected to mandatory AI system documentation, validation evidence, and system characterisation;
- *excludes* organisational processes, conformity assessment workflows, contractual/legal obligations, and post-market monitoring activities;
- does not model complete lifecycle governance, but focuses on regulatory reporting duties;
- refrains from redefining or duplicating content already covered by existing ontologies.

Core Schema Overview

This subsection presents the core schema centred on `aidoc:AISystem`, which is linked to models, datasets, software components, hardware components, agents and documentation artefacts. We describe the main classes and object properties without enumerating all details.

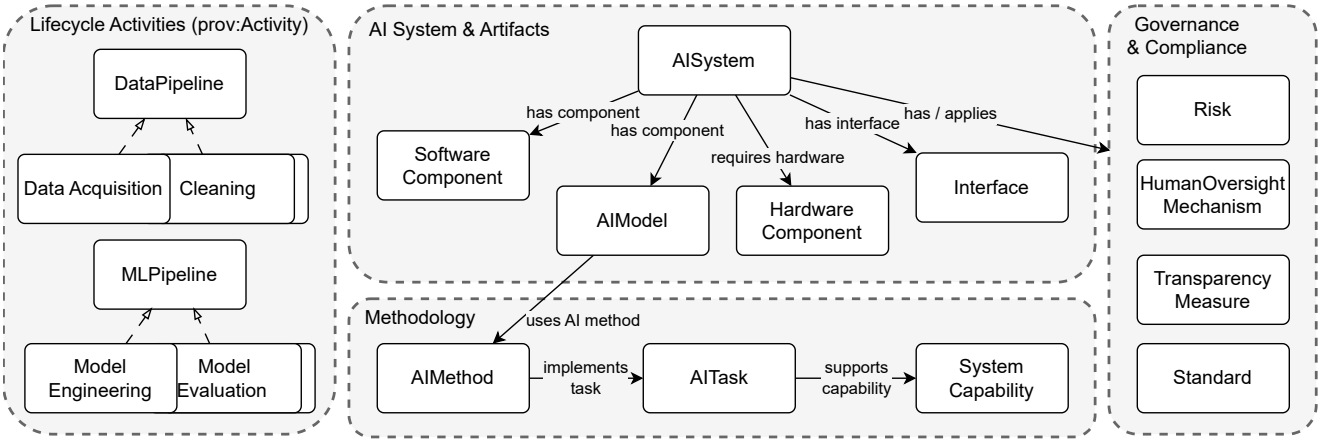


Figure 2. Overview of AIDOC-AP main concepts and relations.

Figure 2 provides a high-level overview of the main concepts: system architecture is modeled through classes such as `aidoc:SoftwareComponent` and `aidoc:AIModel`, linked to `aidoc:AISystem`.

To support structured documentation, AIDOC-AP introduces classes such as `aidoc:DataSheet`, `aidoc:TrainingDataSheet`, `aidoc:TechnicalDocumentation`, `aidoc:ChangeLog` and `aidoc:DeclarationOfConformity`. These artefacts are linked to AI systems and datasets, enabling traceability between raw data, configurations and regulatory documents (not displayed in Figure 2). They differ in their normative status: technical documentation and the declaration of conformity are legal documents required by the AI Act, whereas data sheets represent established community practice for documenting datasets[14], which the AI Act's data documentation requirements (Annex IV(2)(d)) refer to without making them first-class legal documents.

Results: Alignments and Coverage

This section reports the results of the two LLM-assisted pipeline steps: the expert-curated alignment of AIDOC-AP to the reference ontologies, including inter-annotator agreement, the precision of the LLM proposals and a false-negative estimate at the candidate-generation stage, and the evolution of the LLM-estimated Annex IV coverage across ontology iterations, including its run-to-run variance and sensitivity to the sampling temperature.

Reference Ontology Alignments

We provide alignments between AIDOC-AP and several reference ontologies. Alignments are given as RDF graphs specifying SKOS mapping relations such as `skos:exactMatch` and `skos:closeMatch`. The alignment methodology is described in Section Methodology. LLMs are used to propose candidate links based on lexical and semantic similarity; final alignments are manually curated by domain experts. The reported precision scores reflect expert review of LLM-generated proposals and should be understood as quality indicators for the curation process, not as validated inter-ontology mappings. These alignments represent informed suggestions to guide vocabulary reuse, not authoritative semantic equivalences.

Table 4 provides a summary of ontology alignments between AIDOC-AP and reference ontologies, showing the number of alignments per relation type and estimated precision based on manual review of the sampled mappings.

The LLM classification yielded 112 unique suggested correspondences (126 including cross-file duplicates). All three authors independently curated every pair; each judgment is interpreted as a vote for an effective relation (accept = the suggested relation, modify = a corrected relation, reject), and the consensus is the vote supported by a 2-of-3 majority. The 16 pairs without a majority were resolved in a joint adjudication session; both the independent votes and the adjudicated decisions are published in the repository, so the reported agreement is not masked by the adjudication. The curation accepted 44 proposals as suggested, corrected the relation for 14, and rejected 54.

Pairwise inter-annotator agreement (Cohen's κ over accept/modify/reject) ranges from 0.43 to 0.62—moderate agreement that itself illustrates why alignment proposals cannot be adopted unreviewed. The precision of the LLM proposals is 0.39 when requiring the exact suggested relation and 0.52 when counting corrected relations as partial hits. The curation further identified five spurious class-to-property candidates, caused by DPV typing its properties additionally as `skos:Concept`, all of which the LLM had rated `closeMatch`; the entity extraction now excludes property-typed subjects. Human review thus also catches errors upstream of the relation classification.

False negatives can arise earlier in the pipeline, at the lexical candidate-generation step: pairs whose combined lexical similarity falls below the 0.6 operating point are never passed to the LLM. To estimate this miss rate, we curated a seeded random sample of 45 unique pairs of the 146 in the band $[0.5, 0.6)$ just below the operating point, classified with the same LLM step and curated identically: 16 of the 45 pairs were judged meaningful (4 as suggested, 12 with a corrected relation) and 29 were rejected, an estimated false-negative rate of ≈ 0.36 for that band.

Table 4. LLM-proposed candidate mappings and their curation outcome, per target vocabulary. Proposals were generated by Gemma 3 27B (temperature 0) over unique lexical candidates (≥ 0.6) and curated by all three authors (2-of-3 majority vote). Notably every proposal received a self-reported confidence ≥ 0.75, so the confidence score is not a usable discriminator—the expert curation, not the confidence, provides the quality signal.

Vocabulary	Prop.	Acc.	Mod.	Rej.
AIRO	10	8	0	2
VAIR	25	5	2	18
RAINS	9	4	0	5
DPV	8	4	1	3
DPV-AI	21	6	0	15
DPV-AIAct	15	7	2	6
DPV-TECH	6	5	1	0
MLSchema	4	3	0	1
MEX-Core	4	1	1	2
PROV-O	4	0	4	0
Other†	6	1	3	2
Total	**112**	**44**	**14**	**54**

Prop. unique LLM-suggested pairs (cross-file duplicates merged, bucketed by the vocabulary owning the target IRI); Acc./Mod./Rej. curation outcome (accepted as suggested / relation corrected / rejected; majority vote with joint adjudication of split decisions). †DCAT, EP-Plan, FOAF, MLSO, SAO and one syntactically invalid IRI in the VAIR file.

Annex IV Coverage Evolution

We report LLM-estimated coverage scores per requirement across three ontology iterations: (i) initial version with core classes only, (ii) after adding the system architecture module, and (iii) after adding the data requirements module. Table 5 shows the evolution of these estimated scores for the five requirements with the largest improvements.‖

Table 5. Evolution of LLM-estimated coverage scores for the five requirements with the largest improvements across three iterations. Scores averaged across four LLM models (temperature 0, mean over three seeded runs).

Requirement	Iter. 1	Iter. 2	Iter. 3	Gain
Hardware required (Req 5)	0.52	0.88	0.88	+0.35
EU declaration of conformity (Req 21)	0.55	0.62	0.88	+0.33
Illustrations and markings (Req 6)	0.52	0.60	0.81	+0.29
Harmonised standards (Req 20)	0.52	0.57	0.80	+0.28
Software/firmware versioning (Req 3)	0.54	0.80	0.82	+0.27
Average (all 22 req.)	**0.61**	**0.70**	**0.81**	**+0.20**

Note: Iteration 1 contains core classes (`AISystem`, `AIModel`, `Dataset`); Iteration 2 adds system architecture module (`SoftwareComponent`, `ComputationalResource`, `feedsIntoComponent`); Iteration 3 adds data requirements module (`DataSheet`, `LabelingProcedure`, `DataCleaningProcedure`, detailed data properties).

Averaged over all 22 requirements, LLM-estimated coverage increases from 0.61 in iteration 1 to 0.81 in iteration 3, with the largest per-requirement gains for hardware requirements (Req 5, +0.35), the EU declaration of conformity (Req 21, +0.33) and product illustrations and markings (Req 6, +0.29)—requirements addressed precisely by the architecture and documentation modules added in iterations 2 and 3. Table 6 shows that three of the four models (Gemma 3 27B, Llama 3.3 70B, GPT-OSS 120B) improve monotonically and substantially across iterations (+0.22 to +0.35); Apertus 70B is an informative exception, scoring the core ontology already at 0.855 and remaining almost flat (+0.013). We read this not as an absence of improvement but as a reminder that the LLM coverage score is a soft, upward-biased indicator whose *absolute* level is model-dependent—which is precisely why we complement it with expert agreement and the query-based competency-question answering of Section Use Case Validation. Since each requirement is evaluated in an independent single-turn request (Section Methodology), the increase across iterations reflects the ontology modules added between iterations rather than any accumulation of conversational context. Across three seeded runs per model and iteration at temperature 0 the standard deviation of the average coverage score never exceeds 0.009 (mostly 0.000; per-cell figures in the note of Table 6), so for the three models that improve substantially the gains (+0.22 to +0.35) exceed run-to-run noise by more than an order of magnitude. The score is furthermore almost insensitive to the sampling temperature: repeating the whole matrix at temperature 1.0 changes the per-iteration averages by at most 0.013 (0.600/0.695/0.813 vs. 0.613/0.702/0.813 for iterations 1–3).

Table 6. Coverage scores (mean over three seeded runs, temperature 0; run-to-run standard deviations in the note) across ontology iterations for four open-weight LLMs from different model families. Quantisation: Q4_K_M (Gemma, Llama, Apertus), MXFP4 (GPT-OSS). Apertus already scores the core ontology highly and stays nearly flat, whereas the other models rise steeply—consistent with the coverage score being a soft, upward-biased indicator rather than a calibrated metric.

Model	Iter. 1	Iter. 2	Iter. 3	Gain
Gemma 3 27B	0.639	0.764	0.855	+0.216
Llama 3.3 70B	0.564	0.700	0.791	+0.227
GPT-OSS 120B	0.395	0.495	0.740	+0.345
Apertus 70B	0.855	0.850	0.868	+0.013
Average	**0.613**	**0.702**	**0.813**	**+0.200**

Run-to-run standard deviation is 0.000 for all cells (deterministic at temperature 0) except GPT-OSS 120B: 0.001/0.008/0.006 for iterations 1–3.

Use Case Validation

We validate AIDOC-AP in the context of the Horizon Europe project "CERTAIN",** which investigates methods and tools for ensuring compliance, transparency and accountability of AI systems. To assess whether the profile generalises beyond a single system, we construct AIDOC-AP-compliant knowledge graphs for three pilots that deliberately span distinct application domains among the high-risk areas addressed by AI Act Annex III: an

‖Complete experimental results including all requirements, individual LLM outputs, and reproduction scripts are available in the GitHub repository: https://github.com/CERTAIN-Project/aidoc-ap/tree/main/experiments/coverage

**https://certain-project.eu/

energy demand forecasting system for energy communities (critical infrastructure), a *biometric identity verification* system for automated border control (biometrics and border management), and a *retail investment recommendation* system (financial services). The repository additionally contains two further, less complete example instantiations (a civic-participation clustering pipeline and an HR/recruitment assistant), which serve as modelling examples but are not part of the validation. The published knowledge graphs are anonymised and partially fictionalised versions of the pilot documentation (marked as such in the repository); the underlying assessment documents were provided by the developers of the respective pilot systems. The definitive legal classification of the individual pilots under Annex III is performed by the project's legal partners and is not the subject of this paper; here the pilots serve as heterogeneous test cases for the Article 11/Annex IV documentation obligation.

Validation objectives. The validation objectives were: (i) to demonstrate that AIDOC-AP can represent real-world AI system documentation with sufficient granularity, (ii) to evaluate coverage of EU AI Act Annex IV requirements through instantiated knowledge graphs, and (iii) to assess the practical effort required to construct compliant documentation from existing artefacts.

Knowledge graph construction. For each pilot, documentation was gathered via an assessment document filled out by the responsible developers and system engineers together with a system architecture specification. We manually mapped these artefacts to AIDOC-AP classes and properties, creating RDF instances in Turtle format utilizing Protégé[27]. Construction took approximately three person-hours per pilot (ontology engineer working from the system assessment and architecture specification). While this effort is acceptable for high-risk system pilots, future work aims to automate extraction from existing logs (e.g., MLflow) to reduce it.

Table 7. Cross-domain validation across the three instantiated pilot knowledge graphs. "CQ answered" is the number of the 50 Annex IV competency questions that return a bound answer when executed as SPARQL against the graph (see below). Figures are reproducible via `scripts/run_cq_validation.py`.

Pilot	Domain	Tripl.	Ind.	Cls.	CQ
Energy (encom)	Critical infrastructure	374	82	27	38
Biometrics	Biometrics/ border mgmt.	256	46	25	28
Finance (bank)	Financial services	340	70	29	39

"Tripl." total triples, "Ind." individuals, "Cls." distinct AIDOC-AP classes instantiated, "CQ" competency questions answered (of 50).

Cross-domain coverage. Table 7 summarises the three knowledge graphs: they comprise 256–374 triples over 46–82 individuals and instantiate 25–29 distinct AIDOC-AP classes, confirming that the profile represents real documentation with sufficient granularity in three unrelated high-risk domains rather than in a single tailored case.

Empirical competency-question answering. Beyond the LLM-estimated coverage of Section Results: Alignments and Coverage, the instantiated graphs enable a direct, *query-based* coverage measure that is independent of any LLM judgement: we execute all 50 Annex IV competency questions as SPARQL queries (Section Competency Questions) against each graph and count those returning a bound answer. The three pilots answer 28–39 of the 50 CQs (last column of Table 7). Inspecting the unanswered questions is informative: the ones unanswered across *all* graphs are dominated by CQ6.1–6.4 (physical form, external markings, internal layout), which do not apply to software-only AI systems, together with two genuine data-documentation gaps (CQ11.1 training datasheets, CQ11.4 labelling procedures) and one conditional question (CQ20.2, alternative solutions where no harmonised standards are applied). The query-based measure thus both triangulates the LLM-based coverage indicator and surfaces concrete, actionable modelling gaps.

Illustrative examples. Listings 1 and 2 illustrate how AIDOC-AP represents Annex IV documentation across two different domains: the energy pilot (Listing 1) and the biometric border-control pilot (Listing 2). Together they exercise complementary Annex IV facets—general description, lifecycle stages, risks, software and hardware for the former; system components, applied standards, human oversight and transparency measures for the latter—showing that the ontology's expressiveness is sufficient for real-world regulatory documentation in a machine-readable format suitable for automated compliance checking.

```
@prefix aidoc: <https://w3id.org/aidoc-ap#> .
@prefix airo: <https://w3id.org/airo#> .

:Encom a aidoc:AISystem ;
   rdfs:label "Energy Community AI" ;
   aidoc:hasIntendedPurpose :Energy_Forecast ;
   aidoc:hasLifecycleStage
      :Deployment, :Development_Stage, :Operation ;
   aidoc:hasComponent
      :Encom_AI_Model, :Encom_Implementation ;
   aidoc:hasRisk
      :Outside_Data_Manipulation,
      :Unpermitted_predictions ;
   airo:hasRiskControl :User_Training ;
   aidoc:hasSoftwareComponent
      :Admin_Backend, :Apache_Airflow, :Database,
      :Docker, :MLflow, :Minio_Bucket, :Timescale_DB ;
   aidoc:requiresHardware
      :Server_1, :Server_2, :Server_3, :Server_4 ;
   airo:isProvidedBy :Encom_Data_Scientist .

:Energy_Forecast a aidoc:IntendedPurpose ;
   rdfs:label "Predict energy consumption and generation"
   ↪ .
```

Listing 1: Simplified excerpt from the KG: The `AISystem` individual links general system description and intended purposes (Req 1), lifecycle stages (Req 8), identified risks with mitigations (Req 18), software components and hardware requirements (Req 5, Req 10) addressing regulatory requirements via metadata.

Discussion

We extend the NeOn-GPT pipeline with two LLM-assisted steps: (i) alignment to reference ontologies and (ii) iterative evaluation of coverage against competency questions.

These steps reduce manual integration effort and provide systematic, traceable feedback during ontology refinement, which is essential for regulation-driven application profiles.

Interpretation of the results. Read together, our measurements draw a consistent picture. For the alignment step, the LLM proposed 112 unique candidate correspondences across the reference vocabularies, but every proposal carried a self-reported confidence at or above the retention threshold (Section Results: Alignments and Coverage); the confidence is therefore not a usable discriminator, and it is the expert curation—not the model's self-assessment—that provides the quality signal: only about half of the proposals survived curation (precision 0.39 as suggested, 0.52 including corrected relations), and the curators' own pairwise agreement was moderate (κ 0.43–0.62), underscoring that concept alignment requires human deliberation rather than model self-assessment. For the coverage step, three of the four models improve monotonically and substantially across the ontology iterations ($+0.22$ to $+0.35$), with a run-to-run standard deviation below 0.009 and near-identical behaviour at temperature 0 and 1.0; an independent expert assessment of a ten-requirement sample deviates from the pooled scores by 0.12 on average, and every larger deviation is an LLM over-estimate. The fourth model (Apertus) rates even the core ontology highly and stays flat, a reminder that the *absolute* coverage level is model-dependent while the *relative* trend is stable. Finally, the empirical, query-based evaluation answers 28–39 of the 50 competency questions per pilot, independently of any LLM judgement, and its unanswered questions localise both inherent scope boundaries and concrete data-documentation gaps. The LLM-derived scores are thus best read as soft, upward-biased indicators for steering development, whose validity is established not by the models themselves but by human curation and by query-based competency-question answering.

Limitations. Several limitations remain. AIDOC-AP focuses on the technical documentation and lifecycle artefacts of Annex IV, intentionally excluding organisational and legal obligations. The validation covers three CERTAIN pilots across distinct AI Act Annex III high-risk domains, but all are project pilots rather than documentation produced independently by external providers, and broader validation across further domains and providers remains future work. The LLM coverage scores quantify representational adequacy rather than legal compliance; the blind expert comparison covers only a ten-requirement sample, and the query-based competency-question answering checks whether the graph *can* answer a question, not whether the underlying documentation is correct. These indicators should therefore support iterative development, not substitute for legal or expert validation.

Sustainability. Supported by the Horizon Europe "CERTAIN" project (Grant No. 101189650), AIDOC-AP will be actively maintained throughout the project lifecycle: future releases will be driven by validation feedback from seven planned pilot projects, ensuring broad domain coverage. Long-term accessibility is guaranteed via W3ID persistent URIs and a public GitHub repository for issue tracking and version control. Two further aspects support sustainability. First, adaptability: Article 11(3) explicitly anticipates amendments to Annex IV. Because our pipeline treats the requirements as data—formalised requirements with attached competency questions—an amended Annex IV translates into updated CQs, after which the coverage evaluation (Step 7) is re-run to systematically identify the resulting gaps; the methodology is thus a maintenance mechanism, not a one-time engineering exercise. Second, interoperability: by grounding Annex IV-specific terms in DPV concepts where available, AIDOC-AP-based documentation can be consumed by other DPV-based compliance approaches targeting different obligations or regulations. Within the AI Act itself, the same documentation graph could support further obligations, such as the fundamental rights impact assessment by deployers (Article 27) or the information that providers of general-purpose AI models must make available to downstream providers (Article 53); aligning AIDOC-AP's risk concepts with DPV's risk extension[31] is a natural step in that direction. We further intend to contribute Annex IV-specific concepts developed in AIDOC-AP to the DPV community where they add depth beyond the existing extensions.

```
@prefix aidoc: <https://w3id.org/aidoc-ap#> .
@prefix airo: <https://w3id.org/airo#> .

:BiometricRecognitionSystem a aidoc:AISystem ;
   rdfs:label "Biometric Recognition System" ;
   aidoc:hasIntendedPurpose :BorderVerification ;
   aidoc:hasModality :HardwareProduct ;
   aidoc:hasComponent
      :FaceDetectionComponent,
      :PresentationAttackDetectionComponent,
      :BiometricComparisonComponent, :BiometricAIModel ;
   aidoc:appliesStandard
      :ISO_IEC_19795, :ISO_IEC_30107 ;
   aidoc:hasHumanOversight :BorderGuardOversight ;
   aidoc:hasTransparencyMeasure
      :TravellerTransparencyPlan, :BorderGuardTraining ;
   airo:hasAISubject :Travellers ;
   airo:hasAIUser :BorderGuards ;
   aidoc:isProvidedBy :BiometricProvider .

:BiometricAIModel a aidoc:AIModel ;
   aidoc:usesAIMethod :DeepLearningMethod .

:BorderGuardOversight a aidoc:HumanOversightMechanism ;
   rdfs:comment
      "Border guards provide human oversight; in 1:1 mode
      ↪ a match opens the gate." .

:BorderVerification a aidoc:IntendedPurpose ;
   rdfs:label "1:1 identity verification of travellers at
   ↪ automated border gates." .
```

Listing 2: Excerpt from the biometric border-control pilot: the `AISystem` links its intended purpose (Req 1), the software components implementing the recognition pipeline (Req 10), applied harmonised standards (Req 20), human oversight and transparency measures towards travellers and operators (Req 12), together with the affected subjects and users.

Conclusion and Future Work

We presented AIDOC-AP, a resource for representing AI system technical documentation grounded in Annex IV of the EU AI Act. Our approach extends NeOn-GPT with LLM-assisted alignment and coverage evaluation, resulting in an OWL ontology, an Annex IV requirements ontology, a set of competency questions, and alignments to established vocabularies. Three real-world pilot use cases across distinct AI Act high-risk domains, together with an empirical evaluation of competency-question answering, demonstrate the practical applicability of the resource.

Future work includes: (i) extending coverage to additional parts of the EU AI Act beyond Annex IV, (ii) integrating AIDOC-AP with existing audit and monitoring tools (e.g., MLflow) to support continuous assessment, (iii) extending validation beyond the three pilots reported here to further domains and, importantly, to real provider documentation validated by legal experts, (iv) developing a comprehensive library of SPARQL query templates covering the complete set of competency questions to enable automated compliance checking and reporting, and (v) extending the AIDOC-AP framework with SHACL constraint definitions that formally encode Annex IV documentation requirements, enabling automated validation of knowledge graph instances against regulatory obligations.

Acknowledgements

This research was funded by the European Union's Horizon Europe research and innovation programme HORIZON-CL4-2024-DATA-01-01 under grant agreement No. 101189650, and from the Swiss State Secretariat for Education, Research and Innovation (SERI).

Declaration of AI use

Claude Sonnet/Opus was used to partially generate and refine text and tables of this work. Code for the web presentation of data artifacts (ontology, coverage assessment, etc.) was developed utilizing Claude Opus 4.8. See Section Methodology for the use of generative AI in the ontology engineering process.

References

1. Riccardo Albertoni, David Browning, Simon J. D. Cox, Alejandra Gonzalez Beltran, Andrea Perego, and Peter Winstanley. Data catalog vocabulary (DCAT) – version 3. W3c recommendation, World Wide Web Consortium (W3C), August 2024. Latest version: https://www.w3.org/TR/vocab-dcat-3/.
2. Riccardo Albertoni and Antoine Isaac. Introducing the data quality vocabulary (dqv). *Semantic Web*, 12(1):81–97, 2020.
3. Reihaneh Amini, Sanaz Saki Norouzi, Pascal Hitzler, and Reza Amini. Towards complex ontology alignment using large language models, 2024.
4. Muhammad Tuan Amith, Licong Cui, Degui Zhi, Kirk Roberts, Xiaoqian Jiang, Fang Li, Evan Yu, and Cui Tao. Toward a standard formal semantic representation of the model card report. *BMC Bioinformatics*, 23(6):281, 2022.
5. Hamed Babaei Giglou, Jennifer D'Souza, Felix Engel, and Sören Auer. Llms4om: Matching ontologies with large language models. In *The Semantic Web: ESWC 2024 Satellite Events*, pages 25–35, Cham, 2025. Springer Nature Switzerland.
6. Danilo Brajovic, Niclas Renner, Vincent Philipp Goebels, Philipp Wagner, Benjamin Fresz, Martin Biller, Mara Klaeb, Janika Kutz, Jens Neuhuettler, and Marco F. Huber. Model reporting for certifiable ai: A proposal from merging eu regulation into ai development, 2023.
7. Hyojun Choi, Seokju Hwang, and Kyong-Ho Lee. Vspo: Validating semantic pitfalls in ontology via llm-based cq generation, 2025.
8. Ioannis Dasoulas, Duo Yang, and Anastasia Dimou. MLSea: A Semantic Layer for Discoverable Machine Learning. In *The Semantic Web*, 2024.
9. Mike Dean and Guus Schreiber. OWL web ontology language reference. W3C recommendation, W3C, February 2004. Available at https://www.w3.org/TR/owl-ref/.
10. Diego Esteves, Agnieszka Ławrynowicz, Panče Panov, Larisa Soldatova, Tommaso Soru, and Joaquin Vanschoren. Ml schema core specification. Technical report, W3C Machine Learning Schema Community Group, 2016.
11. European Parliament and Council. Regulation (eu) 2024/1689 of the european parliament and of the council of 13 june 2024 on artificial intelligence (ai act). https://eur-lex.europa.eu/eli/reg/2024/1689/oj, 2024. Official Journal of the European Union L 206.
12. European Parliament and Council. Regulation (eu) 2026/1744 of the european parliament and of the council of 8 july 2026 amending regulations (eu) 2024/1689, (eu) 2018/1139 and (eu) 2023/1230 as regards the simplification of the implementation of harmonised rules on artificial intelligence. https://eur-lex.europa.eu/eli/reg/2026/1744/oj, 2026. Official Journal of the European Union, L series, 24 July 2026.
13. Nadeen Fathallah, Arunav Das, Stefano De Giorgis, Andrea Poltronieri, Peter Haase, and Liubov Kovriguina. Neon-gpt: A large language model-powered pipeline for ontology learning. In *The Semantic Web: ESWC 2024 Satellite Events - Hersonissos, Crete, Greece, May 26-30, 2024, Proceedings, Part I*, volume 15344 of *Lecture Notes in Computer Science*, pages 36–50. Springer, 2024.
14. Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Commun. ACM*, 64(12):86–92, November 2021.
15. Gemma Team, Google DeepMind. Gemma 3 technical report, 2025.
16. Delaram Golpayegani, Isabelle Hupont, Cecilia Panigutti, Harshvardhan J. Pandit, Sven Schade, Declan O'Sullivan, and Dave Lewis. Ai cards: Towards an applied framework for machine-readable ai and risk documentation inspired by the eu ai act. In *Privacy Technologies and Policy: 12th Annual Privacy Forum, APF 2024, Karlstad, Sweden, September 4–5, 2024, Proceedings*, page 48–72, Berlin, Heidelberg, 2024. Springer-Verlag.
17. Delaram Golpayegani, Harshvardhan J. Pandit, and Dave Lewis. AIRO: an ontology for representing AI risks based on the proposed EU AI act and ISO risk management standards. In Anastasia Dimou, Sebastian Neumaier, Tassilo Pellegrini, and Sahar Vahdati, editors, *Towards a Knowledge-Aware AI - SEMANTiCS 2022 - Proceedings of the 18th International Conference on Semantic Systems, 13-15 September 2022, Vienna, Austria*, volume 55 of *Studies on the Semantic Web*, pages 51–65. IOS Press, 2022.

18. Delaram Golpayegani, Harshvardhan J. Pandit, and Dave Lewis. To be high-risk, or not to be - semantic specifications and implications of the AI act's high-risk AI applications and harmonised standards. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, FAccT 2023, Chicago, IL, USA, June 12-15, 2023*, pages 905–915. ACM, 2023.
19. Delaram Golpayegani, Harshvardhan J. Pandit, Declan O'Sullivan, and Dave Lewis. Semantic frameworks to support implementation of the EU AI act. OSF Preprint, 2025.
20. Julio Hernandez, Delaram Golpayegani, and Dave Lewis. An open knowledge graph-based approach for mapping concepts and requirements between the eu ai act and international standards. *AI and Ethics*, pages 1–12, 2025.
21. Interoperable Europe (SEMIC), European Commission. MLDCAT-AP: Machine learning DCAT application profile, version 2.0.0. Technical report, European Commission, 2024.
22. Timothy Lebo, Satya Sahoo, and Deborah McGuinness. PROV-O: The PROV ontology. W3c recommendation, World Wide Web Consortium (W3C), April 2013. Latest version: http://www.w3.org/TR/prov-o/.
23. Anna Sofia Lippolis, Mohammad Javad Saeedizade, Robin Keskisärkkä, Aldo Gangemi, Eva Blomqvist, and Andrea Giovanni Nuzzolese. Large language models assisting ontology evaluation, 2025.
24. Llama Team, AI @ Meta. The llama 3 herd of models, 2024.
25. Laura Lucaj, Alex Loosley, Håkan Jonsson, Urs Gasser, and Patrick van der Smagt. TechOps: Technical documentation templates for the AI act, 2025.
26. Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29-31, 2019*, pages 220–229. ACM, 2019.
27. Mark A. Musen. The protégé project: a look back and a look forward. *AI Matters*, 1(4):4–12, 2015.
28. Iman Naja, Milan Markovic, Peter Edwards, and Caitlin Cottrill. A semantic framework to support ai system accountability and audit. In *The Semantic Web: 18th International Conference, ESWC 2021, Virtual Event, June 6–10, 2021, Proceedings*, page 160–176, Berlin, Heidelberg, 2021. Springer-Verlag.
29. Bert Van Nuffelen. DCAT-AP 3.0.1. Application profile specification, SEMIC, July 2025. Previous version: https://semiceu.github.io/DCAT-AP/releases/3.0.0/.
30. OpenAI. gpt-oss-120b & gpt-oss-20b model card, 2025.
31. Harshvardhan J. Pandit, Beatriz Esteves, Georg P. Krog, Paul Ryan, Delaram Golpayegani, and Julian Flake. Data privacy vocabulary (DPV) – version 2. In *The Semantic Web – ISWC 2024*. Springer, 2024. See also the maintained DPV 2.3 release with TECH, AI and AI Act extensions, https://w3id.org/dpv.
32. María Poveda-Villalón, Alba Fernández-Izquierdo, Mariano Fernández-López, and Raúl García-Castro. LOT: An industrial oriented ontology engineering framework. *Engineering Applications of Artificial Intelligence*, 111, 2022.
33. Swiss AI Initiative (EPFL, ETH Zurich, CSCS). Apertus: Democratizing open and compliant LLMs for global language environments, 2025.
34. Mikel Val-Calvo, Mikel Egaña Aranguren, Juan Mulero-Hernández, Ginés Almagro-Hernández, Prashant Deshmukh, José Antonio Bernabé-Díaz, Paola Espinoza-Arias, José Luis Sánchez-Fernández, Juergen Mueller, and Jesualdo Tomás Fernández-Breis. OntoGenix: Leveraging large language models for enhanced ontology engineering from datasets. *Information Processing & Management*, 62(3), 2025.
35. Dawid Wiśniewski. Automatic translation of competency questions into SPARQL-OWL queries. In *Companion Proceedings of The Web Conference 2018 (WWW '18)*, pages 855–859, 2018.
36. Shiyao Zhang, Yuji Dong, Yichuan Zhang, Terry R. Payne, and Jie Zhang. Large language model assissted multi-agent dialogue for ontology alignment. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '24, page 2594–2596, Richland, SC, 2024. International Foundation for Autonomous Agents and Multiagent Systems.