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

Journal Title
XX(X):1–9
©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 LLM-powered ontology engineering framework, 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 in a real-world use case from the CERTAIN project and publish all artifacts under open licenses.

License: CC BY 4.0
DOI: 10.5281/zenodo.17787787
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. Annex IV in particular requires comprehensive information on system architecture, data, training and testing procedures, risk management and lifecycle performance. 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[12], 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, MLSchema, MCRO), including structural similarity scores and curated `skos:closeMatch`, `owl:equivalentClass`, and `owl:equivalentProperty` mappings.
4. **An extended ontology engineering pipeline**, adding two novel LLM-assisted steps, i) semantic alignment and ii) CQ coverage evaluation, to the NeOn-GPT methodology, implemented in an openly available workflow.
5. **A real-world validation** using an operational AI system to produce an application-profile-compliant knowledge graphs.

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 in a real-world use case, Section Discussion discusses limitation

[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

Table 1. Comparison of existing ontologies along five dimensions.

Ontology	System Architecture	Data Req. and Dataset	Lifecycle Coverage	Regulatory Grounding
MLSchema	✓	✓	✓	✗
MCRO	✓	✓	✓	✗
AIRO	✗	✗	✓	✗
VAIR	✓	✗	✓	✓*
DCAT3	✗	✓	✗	✗
PROV-O	✗	✗	✓	✗
AIDOC-AP	✓	✓	✓	✓

*VAIR is motivated by regulatory contexts but is *not* directly grounded in Annex IV.

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[19] and datasheets for datasets[13]. 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[15], VAIR[16], MLSchema[10], MCRO[4], RAINS[21], MLSO[8], DCAT[1], DQV[2], PROV-O[18]. 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 legal or regulatory texts.

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.

Recent work in RegTech and legal informatics has explored modelling regulatory obligations and compliance processes as knowledge graphs[6,14,17]. 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[12] 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[23]. There is also work using LLMs to generate competency questions for validating ontology design, such as identifying semantic modelling pitfalls[7].

However, none of these approaches combine (i) LLM-assisted semantic alignment between ontologies and (ii) LLM-derived coverage estimation for a set of defined competency questions (e.g., from a regulation). To our knowledge, AIDOC-AP is the first pipeline to support both tasks, integrating LLM outputs with automated reasoning and human curation (Section Methodology).

Methodology

Overview of the Extended NeOn-GPT Pipeline

Our ontology engineering methodology is based on the NeOn-GPT pipeline[12], 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 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

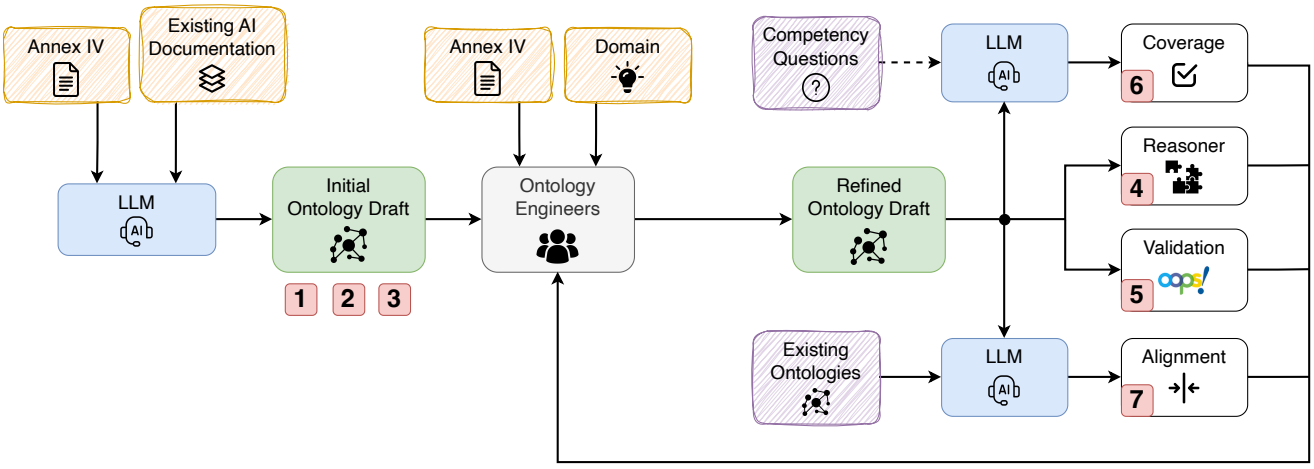

Figure 1. Extended NeOn-GPT pipeline for AIDOC-AP development. The numbers refer to the methodological step described in Section Methodology.

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 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 and VAIR. 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 relation type drawn from a small, fixed set (e.g., `owl:equivalentClass`, `skos:closeMatch`, `skos:broader`, `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). For each candidate pair, we construct a prompt that presents the LLM with the labels, definitions and usage contexts of both terms, and ask it to classify the relationship and assign a confidence score. Only correspondences with p above a fixed threshold (currently 0.75) are retained in the semantic alignment files. Each alignment triple is manually reviewed by domain experts, who can accept, reject or refine the suggested relation.

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:
>
> `owl:equivalentClass`, `skos:closeMatch`, `skos:broader`, `skos:narrower`, or `n/a` (no meaningful relation). Assign a confidence score between 0.0 and 1.0.
>
> Return a JSON object with fields: relation, confidence, explanation.

*The full prompt is available in the project repository.

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 fix temperature to 1.0 for all models to ensure consistency across runs and enable fair comparison between iterations and models.[†]

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. While c_r is not a formally grounded metric and has not been validated against expert assessments, its definition is stable across runs and ontology versions, which allows us to use it as a comparative indicator for iterative ontology refinement. However, this score should be interpreted as decision-support estimates for guiding ontology development, not as compliance guarantees.

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 consolidated text of the EU AI Act, 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 resulting ontology module, published online[§] and as `TTL` file, currently comprises 22 requirement individuals and more than 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

[†] We acknowledge that a lower temperature would reduce sampling stochasticity; future work will explore the impact of temperature on coverage score stability.

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

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

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 2 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[22]: 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 while declaring them *equivalent* to semantically appropriate terms from established vocabularies. 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 imports vocabularies and reuses their semantics through equivalence 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`).

Profile Constraints. The AIDOC-AP term is declared in the `aidoc:` namespace and linked to an external vocabulary term using:

- `owl:equivalentClass`, when AIDOC-AP introduces a class with the same extension as an external class but with context-specific constraints (e.g., cardinality, mandatory documentation roles);
- `owl:equivalentProperty`, when AIDOC-AP introduces a property equivalent in extension to an external property but with narrowed domain, range, or contextual semantics.

Note that `owl:equivalentClass` asserts extensional equivalence, not conceptual identity[9]. This design allows AIDOC-AP to reuse established vocabulary semantics while maintaining its own namespace for specific modeling decisions and documentation requirements.

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 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:VisualDocumentation`, `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).

Results: Alignments and Coverage

Reference Ontology Alignments

We provide alignments between AIDOC-AP and several reference ontologies. Alignments are given as RDF graphs specifying relations such as `skos:closeMatch` and `owl:equivalentClass`. 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

Table 2. 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`

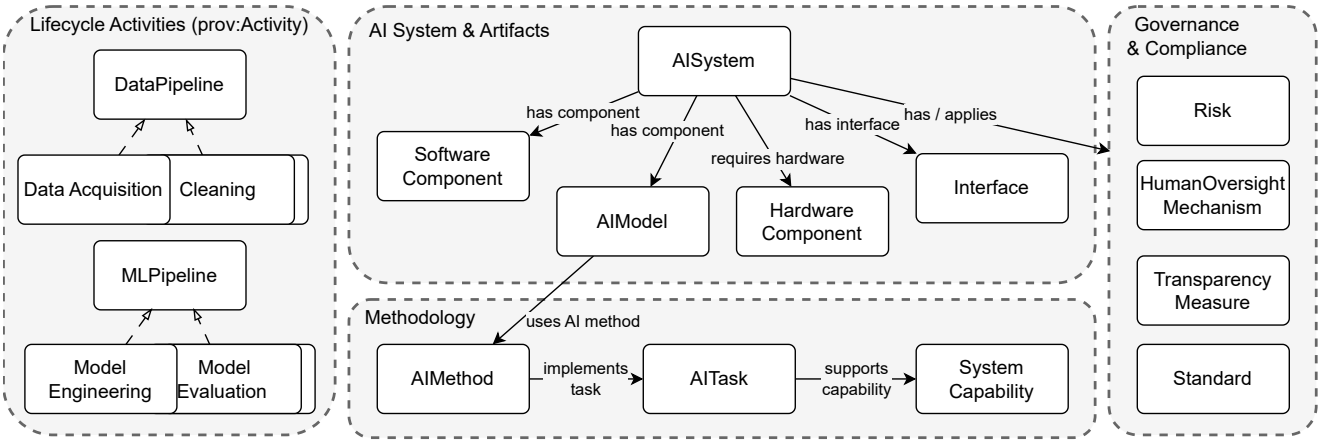

Figure 2. Overview of AIDOC-AP main concepts and relations.

represent informed suggestions to guide vocabulary reuse, not authoritative semantic equivalences.

Table 3 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.

Table 3. "equiv." denotes `owl:equivalentClass` relations; "close" denotes `skos:closeMatch` relations. Precision is based on manual validation, evaluating the appropriateness of relation type. All mappings were generated by an LLM-based alignment bot (Gemma 3:27B) with reported confidence scores ranging from 0.85 to 0.99.

Ontology	equiv.	close	Total	Precision
AIRO	5	1	6	100%
RAINS	2	4	6	66%
VAIR	4	0	4	100%
ML-Onto	2	0	2	100%
MLSchema	2	0	2	100%
MEX-Core	1	0	1	100%
Total	**16**	**5**	**21**	

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 4 shows the evolution of these estimated scores for the five requirements with the largest improvements.[¶]

Table 4. Evolution of LLM-estimated coverage scores for selected requirements with largest improvements across three iterations. Scores averaged across three LLM models.

Requirement	Iter. 1	Iter. 2	Iter. 3	Gain
Declaration of conformity	0.60	0.60	0.95	+0.35
Data governance	0.60	0.95	0.90	+0.30
Technical documentation	0.65	0.75	0.95	+0.30
Cybersecurity	0.60	0.65	0.85	+0.25
Record keeping	0.65	0.85	0.85	+0.20
Average	**0.62**	**0.76**	**0.90**	**+0.28**

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).

[¶]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

On average, LLM-estimated coverage for these high-gain requirements increases from 0.62 in iteration 1 to 0.90 in iteration 3 (45% improvement), with the largest gain for declaration of conformity (Req 21, +0.35), followed by data governance (Req 5, +0.30) and technical documentation (Req 6, +0.30). These trends indicate successful iterative refinement guided by the LLM-based evaluator. Table 5 shows that coverage improvements are consistent across different LLM models (Gemma 3:27B, Llama 3.3:70B, GPT-OSS:120B), with all models showing progressive improvements across iterations.

Table 5. Average coverage scores across ontology iterations for different LLM models.

Model	Iter. 1	Iter. 2	Iter. 3	Gain
gemma3:27b	0.71	0.79	0.86	+0.15
llama3.3:70b	0.66	0.73	0.80	+0.14
gpt-oss:120b	0.48	0.58	0.77	+0.29
Average	**0.62**	**0.70**	**0.81**	**+0.19**

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. We constructed an AIDOC-AP-compliant knowledge graph for an energy demand forecasting system that serves as one of the project's pilot use cases. The system supports several tasks relevant to energy communities, including prediction of energy consumption, prediction of energy generation, and AI-based scheduling of participation factors for metering points.

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. Documentation for the energy demand forecasting system was gathered via an assessment document filled out by the responsible developers and system engineers as well as through a system architecture specification document. We manually mapped these artefacts to AIDOC-AP classes and properties, creating RDF instances in Turtle format utilizing Protégé[20]. The knowledge graph construction process took approximately 10 person-hours for an ontology engineer working on the basis of the system assessment. While 10 hours is acceptable for a high-risk system pilot, future work aims to automate extraction from existing logs (e.g., MLflow) to reduce this effort.

Knowledge graph statistics. Table 6 summarises key statistics of the resulting knowledge graph. The KG comprises 364 triples representing 85 individuals (averaging 4.3 assertions per individual), including 14 software components, 5 datasets, 1 change log, and 9 AI lifecycle activities. The graph instantiates 10 core AIDOC-AP classes, with the most frequently used being `SoftwareComponent` (14 instances) and `SoftwareDependency` (11 instances).

Table 6. Knowledge graph statistics for the CERTAIN energy forecasting use case.

Metric	Count
`AISystem`	1
`AIModel`	1
`Dataset`	5
`SoftwareComponent`	14
`SoftwareDependency`	11
`Interface`	2
`PerformanceMetric`	8
`AIActivity`	9
`ChangeLog`	1
`ModelEvaluation`	1
Total triples	364
Total individuals	85
Doc. artefacts	1
AI activities	9
Vocabularies	5

Illustrative examples. Listings 1 and 2 illustrate how AIDOC-AP enables structured representation of Annex IV requirements. The excerpts show that the ontology's expressiveness is sufficient for real-world regulatory documentation, connecting system components, data flows, and quality measures 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:intendedPurpose
      "Predict energy consumption and generation data" ;
   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 .
```

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.

[‖] https://certain-project.eu/

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

Both steps introduce limitations inherent to LLM-based methods. Outputs may vary across runs and can contain plausible but incorrect assertions. The quality of results remains dependent on training data and prompt design, and generalisability across legal or technical domains is not guaranteed. Importantly, the resulting confidence scores should be interpreted as heuristic indicators for guiding ontology development decisions, not as validated quality metrics.

Limitations. AIDOC-AP focuses on technical documentation and lifecycle artifacts from Annex IV (concerning with technical documentation of high risk AI systems), intentionally excluding organizational or legal requirements. The current validation is limited to a single CERTAIN use case; broader validation across domains is left for future work.

Coverage scores produced by the LLM quantify representational adequacy but do not constitute compliance guarantees. They have not yet been compared against expert assessments or empirical query answering performance. These indicators should therefore be used to support iterative development, not as substitutes for legal or expert validation. In future work we want to include empirical validation against expert annotations and actual competency question answering on instantiated knowledge graphs.

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.

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. A real-world use case demonstrates 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) conducting broader validation across diverse application domains, (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.

```
@prefix aidoc: <https://w3id.org/aidoc-ap#> .
@prefix dqv: <http://www.w3.org/ns/dqv#> .
@prefix dpv: <https://w3id.org/dpv#> .

:Energy_Data a aidoc:Dataset ;
   dcterms:description
      "readings of metering points with quality flags" ;
   dpv:hasDataSource :EEGFaktura ;
   dqv:hasQualityMeasurement
      :Accuracy, :Completeness, :Consistency,
      :Correctness, :Freedom_from_redundancy, :Relevance,
      :Timeliness, :Uniformity .

:Combine_Datasets_Activity a aidoc:DataProcessing ;
   aidoc:hasSourceDataset :Energy_Data, :Weather_Data ;
   aidoc:producesDataset
    :Combined_Energy_and_Weather_Data ;

:Training_Activity a aidoc:DataTraining ;
   aidoc:usesTrainingData
      :Combined_Energy_and_Weather_Data ;
   aidoc:frequency "daily" .
```

Listing 2: Data lifecycle documentation addressing Annex IV data requirements (Req 11): The excerpt demonstrates dataset provenance (`hasDataSource`), quality measurement across several dimensions, data processing workflow combining source datasets, and training activity with daily frequency, supporting documentation of data characteristics and training methodologies.

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

Microsoft Copilot 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 Sonnet. 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. 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.
13. 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.
14. 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.
15. 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.
16. 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.
17. 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.
18. 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/.
19. 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.
20. Mark A. Musen. The protégé project: a look back and a look forward. *AI Matters*, 1(4):4–12, 2015.
21. 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.
22. 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/.
23. 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.