

AEGIS: A Modular Ontology for Governance, Provenance, and Risk in Generative AI Systems

Amna Dridi^{1,*}, Ewan Price¹, Jayanth Chowdary Jaliparth¹, Ramesh Babu Suresh Babu¹, Faiza Javed¹, Priya Mahipal Yadav¹ and Sagor Chandra Sutra Dhar¹

¹ Birmingham City University, Birmingham, United Kingdom

Abstract

Trustworthy deployment of Generative AI (GenAI) requires the ability to trace how models, prompts, outputs, risks, and responsible actors relate to one another. Yet this capability remains fragmented across existing semantic resources, each addressing risk, governance, provenance, or documentation in isolation, and none modelling the complete generation-to-governance chain of an individual GenAI interaction. This paper presents AEGIS, a modular OWL 2 ontology that closes this gap by integrating governance, provenance, validation, and risk within a single, application-oriented semantic model. Developed through a competency-question-driven engineering process, AEGIS explicitly connects AI systems, models, prompts, supporting inputs, generated outputs, identified risks, mitigation strategies, governance roles, and provenance records around individual generation events. We demonstrate AEGIS through a medication-safety review scenario built on a synthetic HL7 FHIR patient record, validating the ontology in Protégé using the HermiT reasoner and confirming logical consistency with no unsatisfiable classes. Functional evaluation against all eight competency questions, translated into SPARQL queries over the instantiated healthcare knowledge graph, successfully retrieved output-level risks, mitigation strategies, responsible actors, and full generation provenance, demonstrating traceable, cross-module retrieval not explicitly supported by the semantic resources examined in this study. While validated here on one high-risk healthcare scenario with manually instantiated individuals, the results establish AEGIS as a compact, extensible semantic foundation for auditable, machine-queryable GenAI governance, with future work extending validation to larger, automatically populated knowledge graphs.

Keywords

Ontology engineering, Generative AI, AI governance, Risk management, Provenance, Trustworthy AI, OWL, SPARQL, Healthcare

1. Introduction

Generative Artificial Intelligence (GenAI), particularly systems based on large language models (LLMs), is increasingly used to support knowledge-intensive tasks such as information synthesis, decision support, content generation, and automated analysis. However, the fluency of generated content does not guarantee its factual correctness or evidential support. Recent research continues to identify hallucination as a significant limitation of LLMs, encompassing outputs that contradict the provided input, previously generated context, or established knowledge [1]. Such failures are particularly consequential in high-impact settings, where unsupported, incomplete, or misleading outputs may influence professional decisions.

The governance of GenAI therefore requires more than the identification of isolated technical risks. It requires mechanisms for documenting the models, prompts, supporting data, generated outputs, evaluations, human actors, and controls involved throughout the AI lifecycle. The NIST Generative AI Profile similarly emphasises the need to govern GenAI risks across system design, development, deployment, use, and evaluation [2]. Recent operational resources also demonstrate the growing need to connect risk definitions with evaluations, benchmarks, mitigation measures, and governance processes [3, 4]. These requirements make traceability and provenance central to the transparent, accountable, and auditable deployment of GenAI systems.

SAGE 2026: Workshop on Semantic AI and Generative AI, 15 September 2026, Ghent, Belgium

*Corresponding author.

✉ amna.dridi@bcu.ac.uk (A. Dridi)

ORCID 0000-0002-0185-103X (A. Dridi)



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

Semantic technologies provide an appropriate foundation for addressing these requirements because they enable heterogeneous technical and governance information to be represented in a machine-interpretable and queryable form. Existing semantic resources have made important contributions to this area. AIRO provides an ontology for representing AI-related risks, impacts, controls, and stakeholders [5], while VAIR supplies more granular terminology for AI risks and controls [6]. The Data Privacy Vocabulary and its AI-related extensions represent technologies, actors, processing activities, risks, measures, and regulatory obligations [7, 8]. PROV-O provides a general semantic model for connecting entities, activities, and agents through provenance relationships [9]. Other resources, including MCRO and AICat, support machine-readable model documentation and AI-system cataloguing [10, 11].

Despite these advances, the relevant capabilities remain distributed across resources with different purposes and levels of abstraction. Risk-oriented ontologies do not generally model the detailed sequence through which a prompt and supporting input are processed by a model to produce a particular output. General provenance ontologies can represent entities, activities, and agents, but do not directly provide GenAI-specific concepts such as prompts, generation logs, output-level hallucination risks, validation findings, and human-review controls. Operational resources such as the AI Risk Atlas and Risk Atlas Nexus connect risks with evaluations and mitigations, but focus primarily on risk resources and application-level governance rather than the traceability of an individual generation event [3, 4]. Consequently, there remains a need for an application-oriented semantic model that connects the complete GenAI generation chain with its governance, risk, validation, and provenance information.

To address this gap, this paper presents **AEGIS**, a modular and application-oriented OWL 2 ontology for modelling governance, provenance, validation, and risk in Generative AI systems. AEGIS represents an explicit GenAI generation process connecting AI systems, models, prompts, supporting inputs, generated outputs, identified risks, mitigation strategies, validation activities, provenance records, and responsible actors. Rather than seeking exhaustive coverage of AI governance, the ontology adopts a competency-question-driven scope intended to remain auditable, implementable, and extensible. AEGIS is also designed to complement, rather than replace, established semantic resources by providing a specialised layer for the governance and traceability of individual GenAI interactions.

The ontology is demonstrated through a medication-safety review scenario based on a synthetic patient record conforming to the HL7 FHIR standard and generated using Synthea [12, 13]. Its quality and functional coverage are evaluated through ontology metrics, logical reasoning, competency-question assessment, and representative SPARQL queries over the instantiated healthcare knowledge graph.

The principal contributions of this work are:

- a modular and application-oriented OWL 2 ontology that integrates GenAI generation, governance, risk management, validation, and provenance within a unified semantic model;
- an explicit representation of individual GenAI generation events, connecting models, prompts, inputs, outputs, risks, mitigations, responsible actors, and provenance records;
- a competency-question-driven development approach and a healthcare instantiation demonstrating the ontology’s applicability in a high-risk context; and
- technical and functional evaluation through logical reasoning, ontology metrics, and SPARQL queries demonstrating traceable and machine-interpretable retrieval across the interconnected ontology components.

The remainder of the paper is organised as follows. Section 2 reviews semantic resources for AI governance, risk, provenance, and transparency. Section 3 describes the ontology engineering methodology. Section 4 presents the AEGIS ontology and its healthcare instantiation. Section 5 reports the technical and functional evaluation, while Section 6 concludes the paper and outlines future research directions.

2. Related Work

Existing semantic resources address complementary aspects of AI governance, including risk management, regulatory compliance, provenance, model documentation, and operational risk cataloguing. Table 1 compares these resources against seven capabilities relevant to the governance of individual GenAI interactions: explicit GenAI concepts, prompts, generated outputs, risks, mitigation mechanisms, governance roles, and provenance. Coverage is classified as *full (F)*, *partial* or *indirect (P)*, or *not explicitly represented (N)*. The classifications were derived from the published specifications and documentation of each resource, based on whether the corresponding concept was explicitly represented, indirectly expressible, or absent.

2.1. AI Risk and Governance Ontologies

AIRO provides a semantic model for AI-related risks, impacts, controls, use cases, and stakeholders, informed by the EU AI Act and established risk-management standards [5]. VAIR complements AIRO with a more granular vocabulary of AI risks, impacts, controls, lifecycle concepts, and forms of human involvement [6]. The Data Privacy Vocabulary and its AI-related extensions broaden this governance perspective by modelling AI technologies, actors, processing activities, risks, mitigation measures, and regulatory obligations [7, 8].

These resources provide strong representations of risk and governance, but they do not explicitly model the complete sequence through which a prompt and supporting input are processed by a model to produce a specific output that is subsequently linked to risks, controls, validation evidence, and responsible actors.

2.2. Provenance, Transparency, and Documentation

PROV-O provides a domain-independent model for representing entities, activities, agents, generation, use, derivation, attribution, and association [9]. MCRO formalises model-card documentation, including intended use, limitations, evaluation results, and model characteristics [10]. AICat extends DCAT to support machine-readable cataloguing of AI systems, models, datasets, providers, and transparency information [11].

Although these resources provide valuable foundations, they operate at different levels of abstraction. PROV-O is intentionally generic, whereas MCRO and AICat focus primarily on static model documentation or catalogue-level metadata. They do not directly represent GenAI-specific elements such as prompts, generation logs, output-level risks, validation findings, and human-review controls within an individual generation event.

2.3. Operational AI Risk Resources

The IBM AI Risk Atlas provides a structured taxonomy of AI risks, including hallucination, harmful content, privacy leakage, explainability, and inappropriate use [3]. Risk Atlas Nexus extends this work by linking risks with models, evaluations, benchmarks, mitigations, and governance resources through an ontology-supported knowledge graph [4]. RO-Crate offers a complementary mechanism for packaging research artefacts and associated metadata [14].

These resources improve operational access to AI-risk knowledge, but their primary focus remains risk cataloguing, evidence packaging, and application-level governance. They provide limited representation of the complete traceability chain of an individual GenAI interaction.

2.4. Research Gap

As summarised in Table 1, existing resources provide strong but fragmented coverage of AI risk, governance, provenance, documentation, and operational risk management. None explicitly integrates the AI

Table 1

Comparison of semantic resources against core capabilities for GenAI governance. F: full coverage; P: partial or indirect coverage; N: no explicit coverage.

Resource	GenAI	Prompt	Output	Risk	Mitigation	Roles	Provenance
AIRO [5]	P	N	N	F	F	F	P
VAIR [15, 6]	P	N	N	F	F	P	N
DPV-AI [7]	P	N	N	F	F	F	P
DPV EU AI Act [8]	N	N	N	F	P	F	P
PROV-O [9]	N	N	N	N	N	P	F
MCRO [10]	N	N	N	P	P	P	P
AICat [11]	P	N	N	N	N	P	P
IBM Risk Atlas/Nexus [3, 4]	P	P	P	F	F	P	P
AEGIS	F	F	F	F	F	F	F

system, model, prompt, supporting input, generated output, output-level risks, mitigations, validation evidence, provenance record, and responsible actors associated with a specific GenAI generation event.

AEGIS addresses this gap through an application-oriented semantic layer that integrates these elements within a single traceable process. It is designed to complement, rather than replace, established resources and to facilitate future semantic alignment with them.

3. Methodology

3.1. Requirement Analysis

Requirements were derived from three sources: AI governance frameworks, the semantic resources reviewed in Section 2, and a healthcare medication-safety scenario. Together, these sources identified the concepts and relationships needed to represent GenAI generation, governance responsibilities, risks, mitigations, validation, and provenance.

3.2. Ontology Engineering Process

AEGIS was developed through an iterative, requirements-driven ontology engineering process comprising six stages, as illustrated in Figure 1: requirements elicitation, conceptual modelling, ontology implementation, scenario instantiation, logical validation, and competency-question evaluation.

Based on these requirements, the identified concepts and relationships were organised into a modular conceptual model and implemented in OWL 2 using Protégé. A medication-safety scenario was then instantiated to assess whether the ontology could represent a realistic GenAI workflow.

The ontology was validated using an OWL reasoner to assess logical consistency and identify modelling issues. Its functional coverage was subsequently evaluated by translating the competency questions into SPARQL queries. Findings from scenario instantiation, reasoning, and query evaluation were used to iteratively refine the ontology before finalisation.

3.3. Ontology Design

AEGIS was designed to provide a machine-interpretable representation of governance, risk, validation, and provenance across the GenAI lifecycle. The ontology is organised into six interconnected modules covering AI systems and models, governance and stakeholders, risks and mitigations, the GenAI

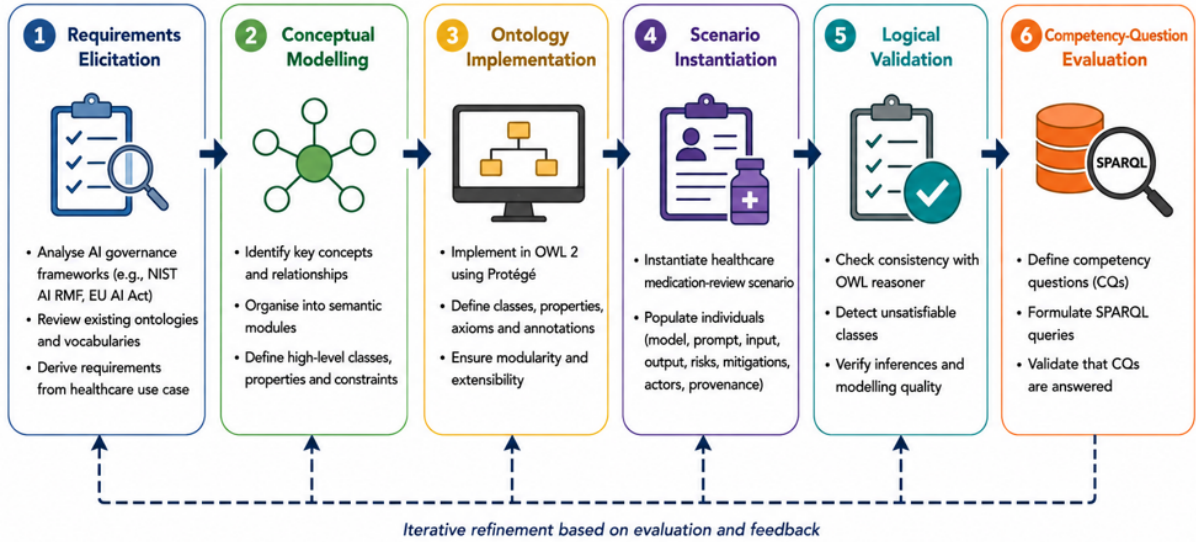


Figure 1: AEGIS ontology engineering workflow.

generation process, validation and monitoring, and provenance. These modules are presented in detail in Figure 2.

The design was guided by four principles: modularity, extensibility, interoperability, and end-to-end traceability. The modular structure separates related areas of knowledge while preserving semantic relationships across them. General classes can be specialised to support new models, application domains, governance frameworks, risks, and mitigation strategies. The ontology was also informed by established resources including AIRO, VAIR, DPV, and PROV-O to facilitate future alignment and reuse [5, 6, 9, 7].

AEGIS was implemented in OWL 2 using Protégé. It uses class hierarchies, object and datatype properties, domain and range axioms, annotations, and selected constraints. This provides sufficient expressiveness for reasoning and semantic querying while avoiding unnecessary logical complexity.

3.4. Competency Questions

Competency questions define the information requirements that an ontology should satisfy and provide a systematic basis for evaluation [16, 17]. The questions listed in Table 2 were derived from the governance, provenance, risk-management, and healthcare requirements identified during requirements analysis. They guided the ontology design and were subsequently evaluated through SPARQL queries in Section 5.

4. The AEGIS Ontology

4.1. Ontology Overview

AEGIS is an OWL 2 ontology designed to provide an integrated semantic representation of governance, risk, validation, and provenance for Generative AI systems. As illustrated previously in Figure 2, the ontology is organised into six interconnected semantic modules.

The current version (v0.3.4) comprises **90 classes**, **37 object properties**, **28 datatype properties**, **34 individuals**, and **304 logical axioms**. The ontology was developed in Protégé and validated using the HermiT reasoner, confirming logical consistency and the absence of unsatisfiable classes. These metrics demonstrate that AEGIS provides a comprehensive yet manageable semantic model for representing GenAI governance.

Table 2
Competency questions guiding the development and evaluation of AEGIS.

CQ	Competency Question
CQ1	Which AI system and model generated a particular output?
CQ2	Which prompt and supporting input were used to generate that output?
CQ3	Which risks are associated with the generated output?
CQ4	Which mitigation strategies address each identified risk?
CQ5	Which actor is responsible for reviewing or approving the generated output?
CQ6	Which provenance record documents the model, prompt, input, output, and user involved in the generation event?
CQ7	Which validation or monitoring activities have been performed for the AI system or generated output?
CQ8	Which governance roles are associated with the AI system and the identified risks?

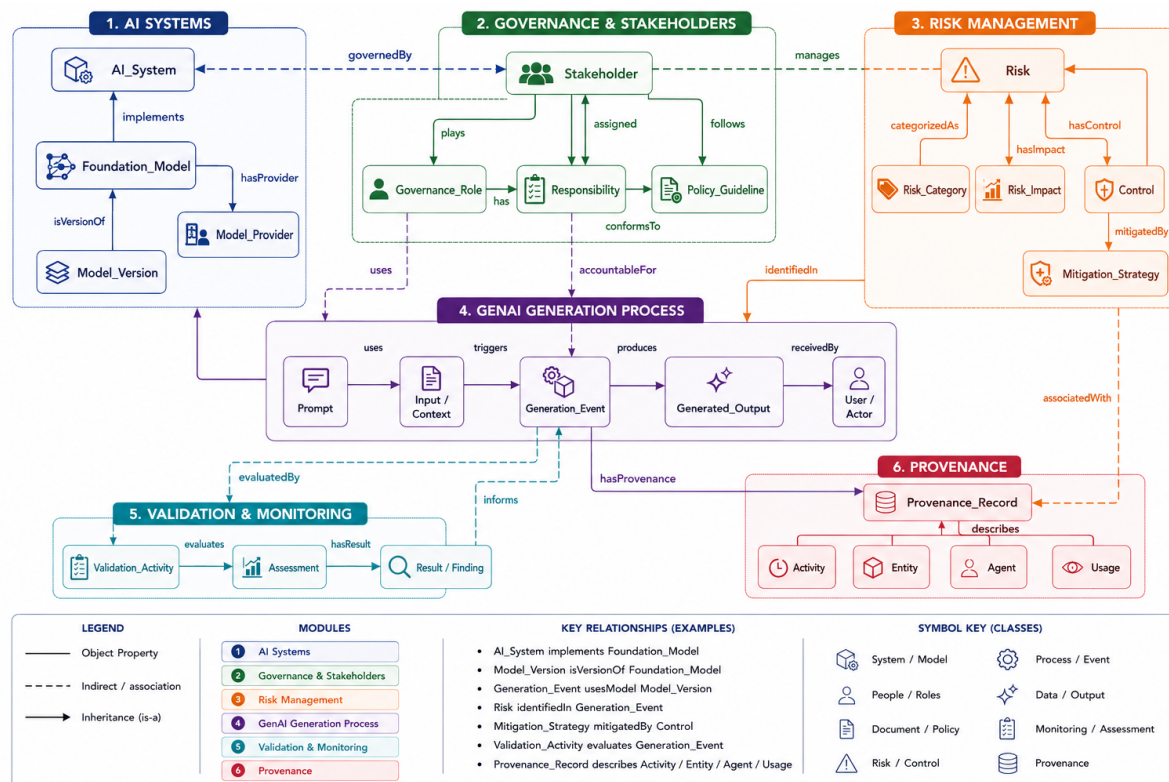


Figure 2: Core classes and relationships represented in the AEGIS ontology. Solid arrows denote direct object-property relations, while dashed arrows indicate broader associations between ontology modules.

AEGIS formally connects these modules through object properties that capture system composition, governance responsibility, risk management, generation activities, validation, and provenance. As illustrated in Figure 2, the GenAI generation process acts as the semantic hub of the ontology, integrating prompts, supporting inputs, AI models, generated outputs, identified risks, mitigation strategies, validation activities, and provenance records into a single traceable workflow.

4.2. Core Concepts and Relationships

Figure 2 presents the principal concepts and semantic relationships represented in AEGIS. The ontology models the complete governance lifecycle of a Generative AI (GenAI) interaction by connecting technical generation processes with governance, risk management, validation, and provenance information. Rather than representing these aspects independently, AEGIS establishes explicit semantic relationships

that enable end-to-end traceability from the initial user input to the final AI-generated output, together with the associated governance artifacts.

Owing to space limitations, this paper presents only the principal concepts and relationships of the ontology. A complete ontology specification, including all classes, object properties, datatype properties, annotations, namespace declarations, and implementation details, is provided in the accompanying Widoco-generated ontology documentation.¹ In addition to these core concepts, AEGIS incorporates specialised classes for AI techniques, hallucination types, evaluation metrics, retrieval mechanisms, compliance requirements, lifecycle stages, audit events, and supporting artefacts, allowing the ontology to represent domain-specific governance scenarios while maintaining a modular structure.

AI Systems and Generation Process. The central component of AEGIS models the GenAI generation workflow. An *AI System* uses one or more *Models* to process a user *Prompt* together with optional supporting *Input* artefacts and produce one or more *Outputs*. These concepts are connected through object properties describing model utilisation, prompt reception, supporting data, and output generation. The generation process therefore acts as the semantic backbone of the ontology, linking the technical execution of GenAI systems with the remaining governance components.

Governance and Risk Management. Governance concepts represent the organisational mechanisms required for the responsible deployment of GenAI systems. The ontology models stakeholders, governance roles, policies, compliance requirements, and decision-making processes, together with their responsibilities. Risk-related concepts capture different categories of GenAI risks, their associated impacts, mitigation strategies, control mechanisms, and responsible actors. These concepts enable risks to be linked directly to generated outputs, facilitating transparent governance, accountability, and regulatory compliance.

Validation and Provenance. The final conceptual group captures validation activities, evaluation metrics, audit information, and provenance records. These concepts document how AI systems and generated outputs are assessed, which evidence supports the resulting decisions, and how individual generation events can be reconstructed. Provenance relationships connect prompts, supporting inputs, models, outputs, risks, mitigations, and responsible actors, thereby providing the traceability required for explanation, auditing, and reproducibility.

Collectively, these conceptual groups provide an integrated semantic representation of GenAI governance in which technical, organisational, and regulatory information are explicitly linked. This unified representation enables complex semantic queries spanning multiple aspects of AI governance and forms the basis for the healthcare instantiation presented in the following subsection.

4.3. Healthcare Use Case Instantiation

To demonstrate the applicability of AEGIS, the ontology was instantiated using a medication safety review scenario derived from a synthetic patient record conforming to the HL7 Fast Healthcare Interoperability Resources (FHIR) standard [12] and generated using the Synthea synthetic patient simulator [13]. Healthcare was selected as the demonstration domain because it represents a high-risk context in which traceability, governance, and accountability are essential for the deployment of Generative AI systems.

Rather than modelling the complete FHIR resource, a representative subset of clinically relevant information was extracted from the synthetic record. This included a patient profile, Type 2 diabetes, chronic kidney disease, Metformin medication, and a serum creatinine laboratory observation. These elements were instantiated as individuals in AEGIS and linked to a medication safety review workflow involving a pharmacist prompt, a clinical review language model, and a generated medication safety summary.

¹<https://zenodo.org/records/21779374>

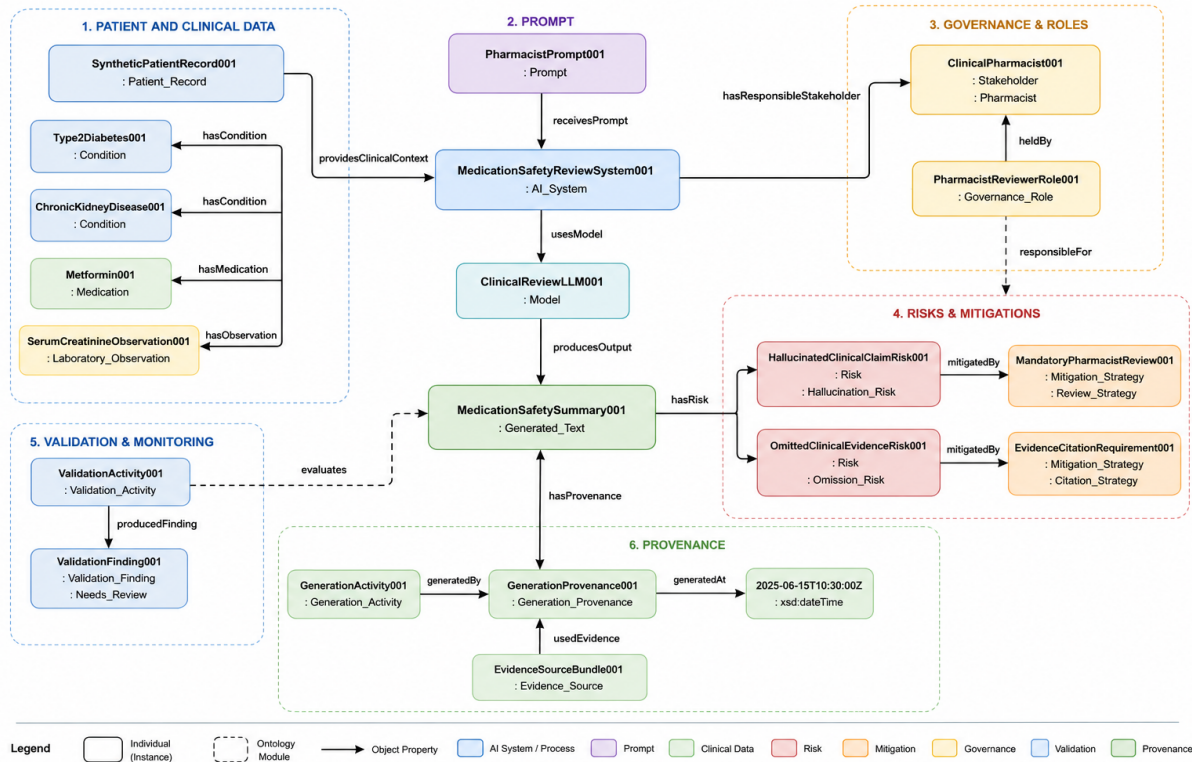


Figure 3: Instance-level representation of the AEGIS healthcare instantiation. A synthetic FHIR-derived patient record and pharmacist prompt are processed by a clinical review model to produce a medication safety summary linked to governance roles, risks, mitigations, validation activities, and provenance information.

Figure 3 presents the resulting instance-level representation. It illustrates how the FHIR-derived clinical information and pharmacist prompt are connected to the AI system and model, and how the generated summary is subsequently linked to governance roles, identified risks, mitigation strategies, validation activities, and provenance information.

As shown in Figure 3, the *MedicationSafetyReviewSystem001* individual receives the *Pharmacist-Prompt001*, uses the *ClinicalReviewLLM001* model, and produces the *MedicationSafetySummary001* output. The system is also linked to a clinical pharmacist and a reviewer role, thereby capturing the responsibility and accountability associated with the generation process.

The generated summary is associated with two representative risks: a hallucinated clinical claim and omitted clinical evidence. These risks are linked to mitigation measures comprising mandatory pharmacist review and an evidence citation requirement. The instantiation therefore demonstrates how AEGIS connects an AI-generated clinical artefact to both the risks it may introduce and the governance controls intended to address them.

The scenario further incorporates validation and provenance information. A validation activity evaluates the generated summary and produces a review finding, while the provenance record captures the generation activity, supporting evidence, and generation time. This enables the complete interaction to be reconstructed and queried across technical, clinical, governance, and accountability dimensions.

The instantiated knowledge graph provides the data basis for the competency-question evaluation and SPARQL queries presented in the following section.

5. Evaluation

The AEGIS ontology was evaluated from both technical and functional perspectives. Technical validation assessed the structural characteristics and logical consistency of the ontology, while functional

Table 3

Summary statistics of the AEGIS ontology.

Metric	Count
Classes	90
Object Properties	37
Datatype Properties	28
Individuals	34
Logical Axioms	304

evaluation examined its ability to satisfy the competency questions and support semantic querying over the instantiated healthcare use case.

5.1. Technical Validation

The technical validation of AEGIS comprised ontology metrics and logical reasoning. Ontology metrics provide an overview of the ontology's size and modelling complexity, while logical validation verifies the consistency of the ontology and the correctness of its axioms. Together, these analyses demonstrate that AEGIS provides a compact yet expressive semantic model suitable for representing Generative AI governance, provenance, risk management, and validation.

5.1.1. Ontology Metrics

Ontology metrics were collected to characterise the structural properties of AEGIS. Table 3 summarises the principal metrics, including the numbers of classes, properties, individuals, and logical axioms. These metrics provide a quantitative overview of the ontology's structural scope and modelling complexity.

5.1.2. Logical Validation

AEGIS was validated in Protégé using the HermiT reasoner. Classification confirmed that the ontology was logically consistent and contained no unsatisfiable classes. The inferred class memberships and relationships within the healthcare instantiation were also inspected and found to be consistent with the intended conceptual model.

5.2. Functional Evaluation

The functional evaluation assessed whether AEGIS satisfies the competency questions defined during the ontology engineering process and supports semantic querying over the instantiated healthcare use case. This evaluation demonstrates that the ontology can retrieve governance, risk, provenance, validation, and AI generation information required to support transparent and accountable Generative AI applications.

5.2.1. Competency Question Evaluation

The eight competency questions defined in Section 3.4 were evaluated using SPARQL queries over the instantiated healthcare knowledge graph. The queries successfully retrieved the required information concerning AI systems, models, prompts, supporting inputs, generated outputs, risks, mitigations, responsible actors, validation activities, governance roles, and provenance records. Owing to space limitations, the following subsection presents two representative queries; the remaining queries were used to verify validation activities and governance-role assignments.

Table 4

Results returned by Query 1.

Output	Risk	Mitigation
MedicationSafetySummary001	HallucinatedClinicalClaimRisk001	MandatoryPharmacistReview
MedicationSafetySummary001	OmittedClinicalEvidenceRisk001	EvidenceCitationRequirement

5.2.2. SPARQL Query Evaluation

Two representative SPARQL queries were executed over the instantiated healthcare knowledge graph. The first retrieves generated outputs, associated risks, and corresponding mitigation strategies. The second retrieves the provenance of a generation event, including the model, prompt, input, output, and responsible user. In the following queries, the prefix `aegis:` denotes the AEGIS ontology namespace: <https://w3id.org/aegis#>. The namespace declaration is omitted from the listings for brevity.

Query 1: Retrieving risks and mitigation strategies. The following query addresses CQ3 and CQ4 by linking each generated output to its identified risks and corresponding mitigation strategies.

Listing 1: Retrieving generated outputs, risks, and mitigation strategies.

```

1 SELECT ?output ?risk ?mitigation
2 WHERE {
3   ?output a aegis:Output ;
4           aegis:hasRisk ?risk .
5   ?risk aegis:mitigatedBy ?mitigation .
6 }
```

The results show that AEGIS links the generated medication-safety summary to two output-level risks and to the governance controls intended to address them.

Query 2: Retrieving generation provenance. The second query addresses CQ1, CQ2, CQ5, and CQ6 by retrieving the generation record and the principal artefacts and actor associated with it.

Listing 2: Retrieving provenance information for a generation event.

```

1 SELECT ?output ?provenance ?model ?prompt ?input ?user
2 WHERE {
3   ?provenance a aegis:Generation_Log ;
4               aegis:recordsOutput ?output ;
5               aegis:recordsModel ?model ;
6               aegis:recordsPrompt ?prompt ;
7               aegis:recordsInput ?input ;
8               aegis:recordsUser ?user .
9 }
```

Table 5

Results returned by Query 2.

Variable	Returned individual
Output	MedicationSafetySummary001
Provenance	GenerationProvenance001
Model	ClinicalReviewLLM
Prompt	SyntheticPharmacistPrompt
Input	SyntheticPatientRecord001
User	ClinicalPharmacist001

The query reconstructs the generation event by identifying the output, provenance record, model,

prompt, supporting patient record, and responsible clinical user. Together, the two queries demonstrate cross-module retrieval across generation, risk, mitigation, governance, and provenance.

5.3. Discussion

The evaluation demonstrates that AEGIS provides a compact yet expressive semantic model for representing key aspects of Generative AI governance, including AI systems, prompts, generated outputs, risks, mitigation strategies, validation activities, and provenance. Logical validation confirmed the consistency of the ontology, while the functional evaluation showed that all competency questions could be answered through SPARQL queries over the instantiated healthcare knowledge graph.

A key strength of AEGIS is its integration of governance, risk management, validation, and provenance within a single ontology. Rather than modelling these dimensions independently, the ontology captures their semantic relationships, enabling transparent tracing of AI-generated outputs from their originating inputs and prompts through to associated risks, mitigation strategies, and provenance records. This integrated representation facilitates explainability, accountability, and semantic interoperability for GenAI applications.

The current evaluation is limited to a single healthcare-inspired use case using manually instantiated individuals. Although this demonstrates the feasibility of the proposed ontology, additional validation across larger datasets and multiple application domains is required to assess its broader generalisability. Furthermore, the ontology currently relies on manual population, and future work will investigate automated population from structured sources such as HL7 FHIR resources and integration with external AI governance and regulatory ontologies.

6. Conclusion

This paper presented AEGIS, a modular and application-oriented OWL 2 ontology for modelling governance, provenance, validation, and risk in Generative AI systems. AEGIS addresses the fragmentation of existing semantic resources by providing an integrated representation of the complete GenAI generation life cycle, connecting AI systems, models, prompts, supporting inputs, generated outputs, governance roles, risks, mitigation strategies, validation activities, and provenance records within a unified semantic framework. The ontology was developed using a competency-question-driven ontology engineering methodology and demonstrated through a healthcare medication safety use case based on synthetic HL7 FHIR data.

The evaluation showed that AEGIS is logically consistent and capable of answering the competency questions through representative SPARQL queries. The results demonstrate that the ontology supports semantic retrieval across governance, risk management, mitigation, validation, and provenance, enabling transparent and traceable representation of GenAI interactions while supporting explainability, accountability, and semantic interoperability.

The current work is limited to a single healthcare-inspired use case using manually instantiated ontology individuals. Although this evaluation demonstrates the feasibility of the proposed approach, further validation across additional application domains and larger knowledge graphs is required to assess its scalability, interoperability, and broader applicability in operational AI governance settings.

Future work will investigate automated ontology population from structured data sources such as HL7 FHIR resources and AI system logs, formal alignment with established semantic resources including AIRO, DPV, VAIR, and PROV-O, and the development of SHACL constraints to support automated ontology validation. We also plan to evaluate AEGIS across additional high-risk domains and evolving AI governance frameworks, including the NIST AI Risk Management Framework and the EU AI Act, to further strengthen its role as an interoperable semantic foundation for trustworthy Generative AI governance.

Data and Artefact Availability

The AEGIS ontology and its accompanying documentation are openly available on Zenodo at <https://doi.org/10.5281/zenodo.21542560>. The archived release contains the OWL ontology, RDF serialisations, example instances, competency queries, provenance information, and generated HTML documentation.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT (OpenAI) for language refinement, improving clarity and readability, and assistance with LaTeX formatting. The authors reviewed and edited all AI-assisted content as needed and take full responsibility for the publication's content.

References

- [1] Y. Zhang, Y. Li, L. Cui, D. Cai, L. Liu, T. Fu, X. Huang, E. Zhao, Y. Zhang, Y. Chen, L. Wang, A. T. Luu, W. Bi, F. Shi, S. Shi, Siren's song in the AI ocean: A survey on hallucination in large language models, *Computational Linguistics* 51 (2025) 1373–1418. doi:10.1162/coli.a.16.
- [2] C. Autio, R. Schwartz, J. Dunietz, S. Jain, M. Stanley, E. Tabassi, P. Hall, K. Roberts, Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile, NIST AI 600-1, National Institute of Standards and Technology, Gaithersburg, MD, 2024. doi:10.6028/NIST.AI.600-1.
- [3] F. Bagehorn, K. Brimijoin, E. M. Daly, J. He, M. Hind, L. Garces-Erice, C. Giblin, I. Giurgiu, J. Martino, R. Nair, D. Piorkowski, A. Rawat, J. Richards, S. Rooney, D. Salwala, S. Tirupathi, P. Urbanetz, K. R. Varshney, I. Vejsbjerg, M. L. Wolf-Bauwens, AI Risk Atlas: Taxonomy and tooling for navigating AI risks and resources, *arXiv preprint arXiv:2503.05780* (2025). doi:10.48550/arXiv.2503.05780.
- [4] I. Vejsbjerg, R. Nair, E. Daly, D. Salwala, S. Tirupathi, Risk atlas nexus: A system for managing AI risks, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, Association for the Advancement of Artificial Intelligence, 2026. AAAI 2026.
- [5] D. Golpayegani, H. J. Pandit, D. Lewis, AIRO: An ontology for representing AI risks based on the proposed EU AI Act and ISO risk management standards, in: *Towards a Knowledge-Aware AI: Proceedings of the 18th International Conference on Semantic Systems*, Studies on the Semantic Web, IOS Press, 2022, pp. 51–65. doi:10.3233/SSW220008.
- [6] D. Golpayegani, H. J. Pandit, D. Lewis, VAIR: Vocabulary of AI risks, 2025. URL: <https://w3id.org/vair>. doi:10.5281/zenodo.10894914.
- [7] H. J. Pandit, B. Esteves, G. P. Krog, P. Ryan, D. Golpayegani, J. Flake, Data privacy vocabulary (DPV)—version 2, *arXiv preprint arXiv:2404.13426* (2024). doi:10.48550/arXiv.2404.13426.
- [8] W3C Data Privacy Vocabularies and Controls Community Group, EU Artificial Intelligence Act Extension, Data Privacy Vocabulary specification, 2026. URL: <https://w3id.org/dpv/legal/eu/aiact>, accessed: 23 July 2026.
- [9] T. Lebo, S. Sahoo, D. McGuinness, PROV-O: The PROV Ontology, W3C Recommendation, World Wide Web Consortium, 2013. URL: <https://www.w3.org/TR/prov-o/>.
- [10] M. Amith, L. Cui, K. Roberts, C. Tao, Toward a standard formal semantic representation of the model card report, *BMC Bioinformatics* 23 (2022) 284. doi:10.1186/s12859-022-04797-6.
- [11] D. Golpayegani, H. J. Pandit, D. Lewis, AICat: An AI cataloguing approach to support the EU AI Act, *arXiv preprint arXiv:2501.04014* (2025). URL: <https://w3id.org/aicat>. doi:10.48550/arXiv.2501.04014.
- [12] Health Level Seven International (HL7), Fhir release 5: Fast healthcare interoperability resources, <https://hl7.org/fhir/>, 2023. Accessed: 24 July 2026.
- [13] J. Walonoski, M. Kramer, W. Nichols, S. Quina, C. Moesel, B. Hall, D. Duffett, A. Dube, T. Gallagher, S. McLachlan, Synthea: An approach, method, and software mechanism for generating

synthetic patients and the synthetic electronic healthcare record, *Journal of the American Medical Informatics Association* 25 (2018) 230–238. doi:10.1093/jamia/ocx079.

- [14] S. Soiland-Reyes, P. Sefton, M. Crosas, L. J. Castro, F. Coppens, J. M. Fernández, D. Garijo, B. Grüning, M. La Rosa, S. Leo, et al., Packaging research artefacts with ro-crate, *Data Science* 5 (2022) 97–138.
- [15] D. Golpayegani, H. J. Pandit, D. 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*, Association for Computing Machinery, New York, NY, USA, 2023. doi:10.1145/3593013.3594050.
- [16] M. Grüninger, M. S. Fox, Methodology for the design and evaluation of ontologies, in: *Workshop on Basic Ontological Issues in Knowledge Sharing*, Montreal, Canada, 1995.
- [17] N. F. Noy, D. L. McGuinness, *Ontology Development 101: A Guide to Creating Your First Ontology*, Technical Report KSL-01-05, Stanford Knowledge Systems Laboratory, 2001.