CertGraph: Towards a Comprehensive Knowledge Graph for Cloud Security Certifications

Stefan Schöberl[1], Christian Banse[2], Verena Geist[1], Immanuel Kunz[2], Martin Pinzger[3]

[1]Software Competence Center Hagenberg GmbH, Hagenberg im Mühlkreis, Austria

[2]Fraunhofer AISEC, Garching bei München, Germany

[3]University of Klagenfurt, Klagenfurt, Austria

stefan.schoeberl@scch.at, banse@aisec.fraunhofer.de, verena.geist@scch.at, kunz@aisec.fraunhofer.de, martin.pinzger@aau.at

Introduction

Using semantic representations in the field of security certifications has recently gained traction, especially through the MEDINA project [4] and related research in this field [2, 1, 3]. These existing approaches build on the notion of gathering so called *evidence* – from sources such as the cloud infrastructure – to demonstrate compliance to certain standards or regulations. To harmonize evidence gathered from various cloud providers and technologies, a mapping to a structure described in an *ontology* is performed.

What is the problem with existing approaches? Previous approaches are not very comprehensive in terms of semantic modelling, for example focusing mostly on cloud infrastructure resources. But, in a real-world certification scenario, many more resource types, such as source code, policy documents or other data assets need to be assessed. Second, previous approaches created different independent evidence for each resource and stored them into *information silos*, even if they describe the same aspect (e.g., configuration of encryption), but from different view points.

How is our approach different? First, CertGraph aims to be a systematic approach to building an ontology for security certifications spanning the *complete stack* from infrastructure layer, source code, data to policies and procedures. Furthermore, it provides an initial approach for the *fusion* of evidence coming from different views/sources of the same resource.

Our Solution

We propose a *knowledge graph-based approach* designed to streamline security certification, in which gathered evidence is transformed to a common format, described in an ontology. Furthermore, multiple gathered evidence describing the same aspect, e.g. the configuration of transport encryption, are fused together.

Ontology design

The *CertGraph Ontology*, see Figure 1, consists of multiple smaller ontologies. Two ontologies form the *base*: *Core* and *Security Feature*. *Security Feature* models security properties and is based on the taxonomy with the same name from the Cloud Property Graph [2]. *Core* models detected or extracted evidence regardless of the actual source.

Extensions are built on top of *Core* and hook into the *Asset* taxonomy. We propose four extensions, each covering its own domain, as shown in Table 1.

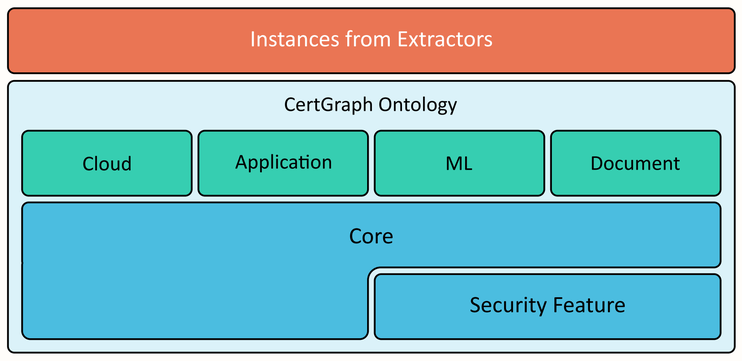


Figure 1: Ontology Architecture

Table 1: Proposed Ontology Domains

Ontology	Description
Cloud	to model cloud resources and this extension is based on the *CloudResource* taxonomy from the Cloud Property Graph [2].
Application	to model source code and code-like artifacts. Here a suitable abstraction level has to be found.
ML	to model machine learning models deployed in the cloud. A suitable starting point could be the Deep Learning taxonomy described by Sarker [5].
Document	to model policy and organizational documents, which primarily contain human-readable text.

Figure 2 shows excerpts of the resulting knowledge graph in detail. Classes (shown as rectangles) and instances (shown as hexagons) for the TLS (Transport Layer Security) example, showing an evidence found in source code (implemented) and a corresponding evidence in an architecture document (specified) regarding transport encryption, which can be used to verify CRY-02 from BSI C5:2020.

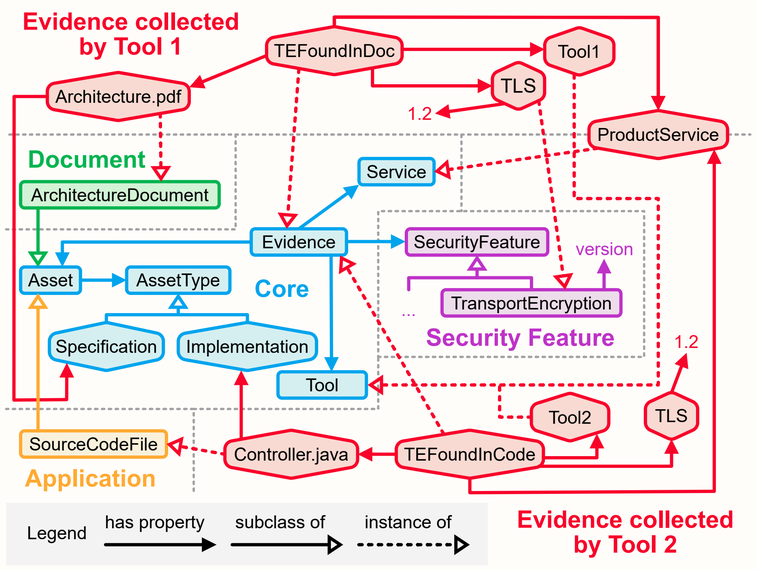


Figure 2: Excerpt of the knowledge graph including instances

Each *Evidence* is connected to a *SecurityFeature*, to a *Tool* (to link the extractor for traceability), to an *Asset* (to store the detection point for traceability) and to a *Service* (to link to the related cloud service).

Asset has a connection to *AssetType* (modeled as an enumeration type), to distinguish between specified and implemented behaviour.

Approaches for knowledge fusion

SWRL or similar languages to describe rules, which are used to derive new knowledge from gathered evidence, thus new edges are added to the graph, which in turn leads denser interlinking of data.

Unique IDs are necessary to identify service instances (i.e. each service can be referenced by a unique URI across extractors).

SPARQL to query the graph and in this way to link the information in the graph and receive it as a query result. Currently we are evaluating, what can be implemented, which libraries are available and what is supported by the used graph database.

Potential usage and benefits

- Evidence collected from different extractors is not considered in isolation from one another.
- No big *All-In-One* extractor, but rather multiple smaller specialized ones.
- Extractors are designed to complement each other in their respective domain, which also improves the reliability of the outcome.
- Gathering and creating the information basis for certification is mostly automated, which otherwise would be a time-consuming task.

Conclusions

Summary: We provided a systematic approach for building a comprehensive ontology for cloud security certifications as well as initial ideas for knowledge fusion. We encourage other collaborations with other domains that can be included in the ontology.

Future work: Next steps include

- further formalization of concepts within the extensions,
- evaluation, which formalism is supported by libraries and databases, and
- modeling and implementation of knowledge fusion in software.

Acknowledgements: Funded by the EU project EMERALD, Grant No. 101120688.

References

[1] Christian Banse, Immanuel Kunz, Nico Haas, and Angelika Schneider. 2023. A semantic evidence-based approach to continuous cloud service certification. In *Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing*, 24–33.

[2] Christian Banse, Immanuel Kunz, Angelika Schneider, and Konrad Weiss. 2021. Cloud property graph: connecting cloud security assessments with static code analysis. In *2021 IEEE 14th International Conference on Cloud Computing (CLOUD)*.

[3] Immanuel Kunz, Konrad Weiss, Angelika Schneider, and Christian Banse. 2023. Privacy property graph: towards automated privacy threat modeling via static graph-based analysis. *Proceedings on Privacy Enhancing Technologies*.

[4] Leire Orue-Echevarria, Juncal Alonso, et al. 2021. Medina: improving cloud services trustworthiness through continuous audit-based certification. In *CEUR Workshop Proceedings*. CEUR-WS.

[5] Iqbal H Sarker. 2021. Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions. *SN computer science*, 2, 6.