Assessment and Management of Organisational Evidences - AMOE

Franz Berger

MEDINA project - Fabasoft

Objectives

The AMOE component is focusing on providing a solution to enable continuous and semi-automatic auditing for cloud service providers using the MEDINA framework.

Introduction

Cloud security schemes like the ENISA EUCS [1] include some controls and requirements that are of organisational nature, meaning they are not suitable to be automatically monitored like technical requirements. Therefore, a subtask of the MEDINA project has been dedicated to find a solution. The result of the related research activities is a simple prototype based on pretrained models. As MEDINA's assessment strategy for technical requirements is based on metrics the approach was adapted to organisational requirements resulting in the usage of so called organisational metrics. Each org. metric consists of some keywords that are used to reduce the search space. The metric description is a simple question targeted to measure a specific feature commonly found in text documents relevant for audits. The metric target value is used for assessment hints to be provided to the compliance manager (user). The compliance manager or auditor needs to confirm that the extracted evidence (aided by an assessment hint if possible) fullfills the complieance status to an org. metric. This can be done in an interactive GUI that displays the processed document as well as the original for double checking the information. Given well defined metrics, this could speed up the auditing process. The policy data processed is sensitive for the security of a cloud service and thus big datasets are being difficult to obtain. The relevant evidence parts need to be annotated for further training of new models and quality measurements of the prototype.

Data

The experiments and research conducted is based on unstructured textual policy data. The queries for the evidence extraction is based on the organisational metrics which are created specificly for the MEDINA project. An example would be:

- **metric name:** LogDataRetentionTimeQ1
- **description:** How long is log data stored?
- **keywords:** logging, monitoring
- **scale:** days
- **operator:** <=
- **target value:** 100
- **data type:** int

Every security requirement of the EUCS is linked to multiple metrics that can be used to assess concrete parts defined in the rather generic requirements. Depending on the Cloud Service Provider (CSP) and number of could services covered, policy documents can be very long (gt. 50 pages). Therefore the input data needs to be reduced to speed up processing time, which can also lead to more precise results.

Evidence extraction

The extraction of evidence snippets is based on a pre-trained question answering (QA) system (roberta-base-squad for QA[2]). The bottom part of Figure 1 shows the pipeline steps for evidence extraction. First, the input text document is filtered using the org. metric keywords to reduce search space and thus processing time for the QA model. Then the selected sections are used to query the potentially relevant evidence text using the metric description (question). If the org. metric has set a target value - the output of the model is translated into similar format (if possible) and an assessment hint is computed by checking the output against the target value with the defined metric operator. The QA model provides a score that could aid in determination of whether the output is relevant (not all queries produce relevant output). However, here is a promising research result for the example metric listed in the poster's data section (extracted answer in bold): "How long is log data stored?" answer of QA: "From a operational necessity standpoint, we therefore configure the log retention time to a maximum of **90 days** after which log data are automatically deleted."

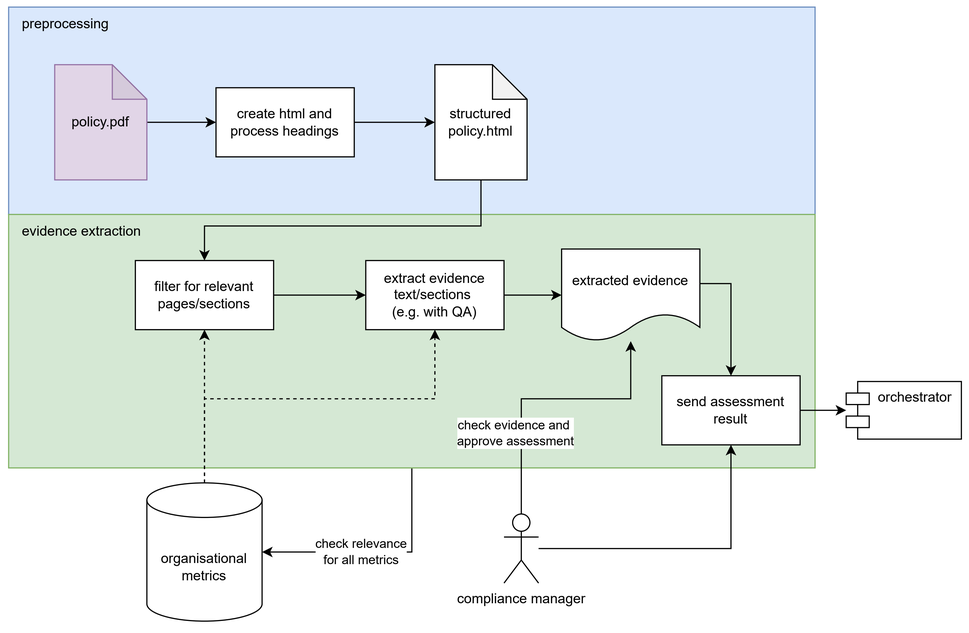

Figure 1:Architectual overview of the prototype

Pre-processing

MEDINA research discussion have shown that most of the CSP's policy documents are available as unstructured text documents (e.g. PDFs). To retain some of the structure given by e.g. section headings, the PDF documents are pre-processed to HTML. Depending on the document origin. some headings need further rule based recognition. This process is depicted in the upper part of Figure 1.

Conclusion

The tool presented here could be an useful extension for any CSP to automate the auditing of textual policy documents. To make this tool future-proof, however, the challenge of creating a suitable dataset remains. In the future, other approaches such as text similarity can also be incorporated as well as further research on pre-processing.

References

[1] ENISA EUCS – Cloud Services Scheme. `https://www.enisa.europa.eu/publications/eucs-cloud-service-scheme`.

[2] Question answering model - roberta-base-squad2. `https://huggingface.co/deepset/roberta-base-squad2`.

Contact Information

- Web: https://medina-project.eu/
- Email: franz.berger@fabasoft.com

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No: 952633.