Exploring GDPR Compliance Over Provenance Graphs Using SHACL

Harshvardhan J. Pandit, Declan O'Sullivan, and Dave Lewis

ADAPT Centre, Trinity College Dublin, Dublin, Ireland
{harshvardhan.pandit|declan.osullivan|dave.lewis}@adaptcentre.ie

Abstract. Semantic web technologies provide an open and adaptable framework for compliance regarding the General Data Protection Regulation (GDPR). Our previous work in this regard demonstrates the use of SPARQL for querying provenance of consent and personal data lifecycles for compliance. We extend this work through our model for evaluation of GDPR compliance using SHACL to validate the correctness and completeness of information. The model describes the creation of a compliance graph consisting of information required to document and demonstrate compliance linked to specific articles and obligations within the GDPR using the GDPRtEXT vocabulary.

Keywords: GDPR · Regulatory Compliance · Provenance · SHACL

CEUR-WS.org/Vol-2198/paper_120.pdf

1 Introduction

Provenance is one of the important categories of information regarding compliance with the General Data Protection Regulation (GDPR) due to obligations surrounding how consent and personal data are collected, stored, used, and shared. Semantic web technologies have been proven to provide an open and extensible framework for representation and querying of information related to such obligations [1,2,4,5]. Our previous work [5] in this regard demonstrated the modelling and querying of provenance information related to GDPR compliance obligations using semantic web technologies. It provided a proof-of-concept demonstration[1] for the querying of compliance-related information using SPARQL queries based on the GDPR readiness checklist published by Ireland's Data Protection Commissioner's office[2].

We extend this work through our model for evaluation of GDPR compliance using SHACL[3] to validate the correctness and completeness of information. The model describes the creation of a compliance graph consisting of information required to document and demonstrate compliance linked to specific articles and obligations within the GDPR using the GDPRtEXT [3] vocabulary.

[1] https://w3id.org/GDPRep/checklist-demo
[2] http://gdprandyou.ie/
[3] https://www.w3.org/TR/shacl/

2 Related Work

The SPECIAL consent, transparency and compliance framework [2] defines RDF vocabularies representing data subject's consent as usage policies and data processing and sharing events as provenance logs. It performs GDPR compliance checking using OWL reasoning and uses a modular architecture that demonstrates the feasibility of semantic web based approaches for GDPR compliance. Agarwal Et al. [1] extend ODRL to represent GDPR obligations in their compliance assessment tool where the obligations are linked to their relevant articles in GDPR. The work described in this paper takes a similar approach with the major differing point being the focus on provenance and use of SHACL for validation.

3 Validation Model

Our model for GDPR compliance, as depicted in Fig. 1, consists of three parts - querying (covered in previous work), validating retrieved information (described in this paper), and generating documentation (planned future work). The provenance information is represented using the GDPRov vocabulary and is linked to concepts within GDPR using the GDPRtEXT vocabulary. The model is further explained with an example use-case in an online article[4].

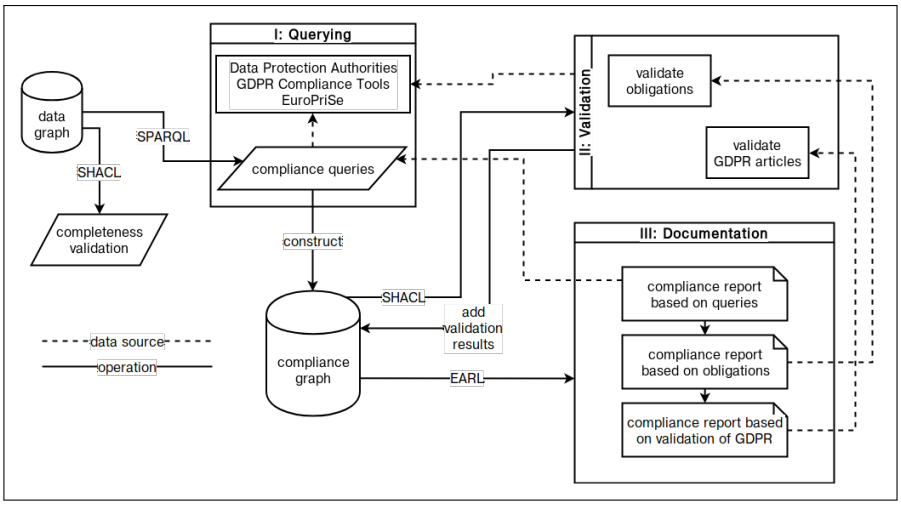

Fig. 1. Model for evaluation of GDPR compliance using SHACL for validation

The feasibility of validation on individual data instances at large scale is questionable due to complexity of analysis. We currently focus only on the model of the system as an abstract representation of the processes and artefacts and their

[4] http://openscience.adaptcentre.ie/projects/CDMM/compliance/model.html

interactions. Compliance is therefore an assessment of the organisation's data practices rather than an investigation of a particular data subject's activities.

The requirements for creating validation shapes is gathered from an analysis of the GDPR legal document, various informative articles published by Data Protection Authorities and commercial organisations, and auditing organisations such as European Privacy Seal (EuroPriSe). These are then used to create a set of questions (similar to those within the GDPR readiness checklist) which retrieve information associated with compliance. The questions are then expressed as SPARQL queries and can be adapted for creating shapes for compliance validation using the SHACL-SPARQL extension.

Validation is performed using SHACL (Shapes Constraint Language), a language for validating RDF graphs against a set of conditions termed as shapes and expressed as RDF graphs themselves. The validation occurs in two distinct stages - first stage (completeness) ensures presence of required information and is carried out before any compliance queries are executed. Ideally, the completeness of the graph is maintained in a continuous fashion. The second stage (compliance) checks for conformance to specific obligations to evaluate compliance. The process of validation is captured using PROV-O to record its execution.

Compliance validation takes place on a separate graph, termed compliance graph, which is constructed from queries that retrieve and structure the required information. The information within the compliance graph is linked to relevant obligations within the text of the GDPR using the GDPRtEXT ontology. Compliance validation is then performed using SHACL and the assessment is added to the graph. The purpose of this graph is to represent information relevant to compliance, to keep it separate from the data graph which may change with time, to capture a snapshot of the state of compliance, and to assist in the generation of compliance documentation.

The compliance validation process occurs in two stages. In the first, shapes are validated and results are linked to specific obligations and added to compliance graph. The second stage of validation tests outcomes of the first stage against their linked obligations to allow for reuse of shapes to test different obligations, and to generate validation directly linked to GDPR articles based on fulfilled obligations. At the end of the compliance validation process, the compliance graph contains information required to answer the compliance queries, their results linked to specific obligations within GDPR, and an indication of the compliance status for GDPR articles.

Documentation and demonstration of compliance can then be performed as SPARQL queries on the compliance graph, and persisted as a compliance report using the EARL vocabulary. Further investigation of how and why the compliance status was achieved is possible by exploring the validation reports and queries present in the compliance graph. This can be exploited to create a tool for top-down exploration of compliance that can be used to document and demonstrate compliance in an interactive fashion.

4 Conclusion & Future Work

This paper extends our our previous work [5] on querying compliance-related information into a model for evaluating GDPR compliance based on provenance of consent and personal data lifecycles using the validation mechanism provided by SHACL. The approach consists of creating a compliance graph consisting of information required for answering compliance-related queries as well as results for conformance to various GDPR obligations. It uses the GDPRtEXT vocabulary to link validation results to specific concepts and articles within GDPR which allows the creation of interactive compliance documentation. The model is further described through an example use-case in an online article[5].

In terms of future work, we plan to create a proof-of-concept demonstration of using SHACL to test compliance obligations with a focus on interactive documentation as described in this paper. There is also the possibility of using graph reduction and summarising techniques to simplify the validation process by representing legal obligations as patterns and testing them for compliance. Finally, we plan to explore compliance coverage as a measure to compare our work with similar work.

References

1. Agarwal, S., Steyskal, S., Antunovic, F., Kirrane, S.: Legislative Compliance Assessment: Framework, Model and GDPR Instantiation. Annual Privacy Forum (APF 2018) (in-press) (2018)
2. Kirrane, S., Fernández, J.D., Dullaert, W., Milosevic, U., Polleres, A., Bonatti, P., Wenning, R., Drozd, O., Raschke, P.: A Scalable Consent, Transparency and Compliance Architecture. In: Proceedings of the Posters and Demos Track of the Extended Semantic Web Conference (ESWC 2018) (in-press) (2018)
3. Pandit, H.J., Fatema, K., O'Sullivan, D., Lewis, D.: GDPRtEXT - GDPR as a Linked Data Resource. In: The Semantic Web. Lecture Notes in Computer Science, Springer, Cham (Jun 2018). https://doi.org/10.1007/978-3-319-93417-4_31
4. Pandit, H.J., Lewis, D.: Modelling Provenance for GDPR Compliance using Linked Open Data Vocabularies. In: Proceedings of the 5th Workshop on Society, Privacy and the Semantic Web - Policy and Technology (PrivOn2017) (PrivOn) (2017), http://ceur-ws.org/Vol-1951/#paper-06
5. Pandit, H.J., O'Sullivan, D., Lewis, D.: Queryable Provenance Metadata For GDPR Compliance. In: SEMANTiCS 2018 – 14th International Conference on Semantic Systems (in-press). Vienna, Austria (2018), https://s3-eu-west-1.amazonaws.com/harshp-media/research/publications/2018_conference_queryable_provenance_metadata_for_GDPR_compliance.pdf

Acknowledgements

This work is supported by the ADAPT Centre for Digital Content Technology which is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund.

[5] see footnote 4