Noname manuscript No.
(will be inserted by the editor)

Modeling Data Protection and Privacy: Application and Experience with GDPR

Damiano Torre [3 1] · **Mauricio Alferez** [1] · **Ghanem Soltana** [1] · **Mehrdad Sabetzadeh** [2 1] · **Lionel Briand** [1 2]

Received: date / Accepted: date

Abstract In Europe and indeed worldwide, the General Data Protection Regulation (GDPR) provides protection to individuals regarding their personal data in the face of new technological developments. GDPR is widely viewed as the benchmark for data protection and privacy regulations that harmonizes data privacy laws across Europe. Although the GDPR is highly beneficial to individuals, it presents significant challenges for organizations monitoring or storing personal information. Since there is currently no automated solution with broad industrial applicability, organizations have no choice but to carry out expensive manual audits to ensure GDPR compliance. In this paper, we present a complete GDPR UML model as a first step towards designing automated methods for checking GDPR compliance. Given that the practical application of the GDPR is influenced by national laws of the EU Member States, we suggest a two-tiered description of the GDPR, generic and specialized. In this paper, we provide (1) the GDPR conceptual model we developed with complete traceability from its classes to the GDPR, (2) a glossary to help understand the model, (3) the plain-English description of 35 compliance rules derived from GDPR along with their encoding in OCL, and (4) the set of 20 variations points derived from GDPR to specialize the generic model. We further present the challenges we faced in our modeling endeavor, the lessons we learned from it, and future directions for research.

Keywords General Data Protection Regulation (GDPR) · Conceptual Modeling · Model Variability · Regulatory Compliance · Unified Modeling Language (UML)

Damiano Torre
E-mail: damiano.torre@tamuct.edu

Mauricio Alferez
E-mail: alferez.mauricio@gmail.com

Ghanem Soltana
E-mail: ghanem.soltana@gmail.com

Mehrdad Sabetzadeh
E-mail: m.sabetzadeh@uottawa.ca

Lionel Briand
E-mail: lionel.briand@uni.lu

[1] SnT Centre for Security, Reliability and Trust, University of Luxembourg, Luxembourg

[2] School of Electrical Engineering and Computer Science, University of Ottawa, Canada

[3] Department of Computer Information Systems, Texas A&M University Central Texas, United States

1 Introduction

With the growing concerns about data protection and privacy, it is becoming increasingly important to assess compliance with the relevant regulations. In Europe and indeed worldwide, the General Data Protection Regulation (GDPR) [14] is now widely viewed as a benchmark for data protection and privacy regulations. The GDPR came into effect in May 2018, replacing the previous Data Protection Directive, 95/46/EC. The GDPR has been designed to harmonize data privacy laws across Europe in order to provide further protection and capabilities to individuals for controlling their personal data in the face of new technological developments [13]. While undoubtedly beneficial to individuals in many ways, the reality with the GDPR is that organizations are having severe difficulties in understanding

what compliance means in this new environment and how to implement the GDPR [37].

In order to comply with the requirements of the GDPR, organizations need to consider the principles of personal data processing as set out in the GDPR and to make regular reviews of their measures, practices and processes regarding the collection, use and protection of personal data. Failure to comply with the GDPR may result in fines of up to 20m or 4% of an organization's global turnover for specific breaches [14]. In addition, organizations are liable for damages and other remedies towards individuals in case of data breaches [15]. For this reason, there is now a fast-growing need for cost-effective methods that will help different business sectors achieve, demonstrate and maintain compliance with the GDPR. Given the sheer complexity of the systems and services that are subject to the GDPR, e.g., e-Government applications and cloud-based services, automated support for GDPR analysis is critically important. At the moment, there is a lack of such support on the market. This gap will become even more evident once individuals start to exercise their rights under the GDPR, likely resulting in an onslaught of new legal challenges for companies. Due to the absence of automated solutions, we have started a long-term investigation, involving both IT researchers and legal experts, into GDPR compliance automation. Our ultimate goal is to create opportunities for developing innovative GDPR-related services. The purpose of the model-based representation of the GDPR (both UML diagrams and OCL constraints) is to help: (a) providing structured knowledge about the terminology that underlies GDPR. This makes our model a useful instrument for communication between different stakeholders, including non-technical ones such as legal experts and software engineers, and (b) developing future automated methods for assessing GDPR compliance. For example, in another recent work [39], starting from the work we have done in this paper, (i.e., considering the identification of the interdependencies between legal basis and data subject rights), we developed an automated AI-based method for checking whether a given privacy policy complies with the provisions of GDPR.

The GDPR is considered the most far-reaching and technically demanding personal data privacy regulation ever established. The high level of rigor that ensuring GDPR compliance entails is increasingly comparable to what is required for demonstrating compliance to safety standards and regulations. GDPR compliance analysis can thus benefit from existing work where models have been employed for systematic compliance analysis in the context of safety certification [26]. While highly advantageous, encoding the GDPR and its compliance mechanisms into a model-based representation is a complicated task. In particular, the level of abstraction of such a representation has to be suitable for ensuring a consistent implementation and interpretation of the regulation, national laws and case law.

In this paper, we draw on Model-Driven Engineering (MDE) [4] for building a machine-analyzable representation of the GDPR as a first step towards the development of future automated methods for assessing GDPR compliance. Although MDE is primarily a paradigm for reducing the complexity of systems development [16], over the years, MDE has outgrown its traditional use and is now increasingly applied as a general mechanism for structuring domain knowledge. When employed in this broader sense, as we do in our work, MDE provides an effective communication bridge between IT experts and domain experts, such as legal experts, who may have little or no software development expertise.

What we pursue in this paper through the application of MDE is a visual and yet precise representation of the textual content of the GDPR. Since a concrete implementation of the GDPR is affected by the national laws of the EU member states, the GDPR's expanding body of case law and other contextual factors, we propose a two-tiered representation of the GDPR: a generic tier and a specialized tier. The generic tier captures the concepts and principles of the GDPR that apply to all contexts, whereas the specialized tier describes a specific tailoring of the generic tier to a given context, including the contextual variations that may impact the interpretation and application of the GDPR. We represent both the generic and specialized tiers using UML class diagrams [25] and a set of invariants expressed in the Object Constraint Language (OCL) [24]. In particular, as we explain in detail in Section 3, we provide an overview of our long-term research project involving four steps: (1) building a generic tier of the GDPR, (2) tailoring the generic tier into a specialized one, (3) developing tool support for representing technical and legal documents in a structured form, and (4) enabling checking GDPR compliance. In this paper, we focus exclusively on conducting steps 1 and 2; the initial results of steps 3 and 4 are presented elsewhere in recently published work [39].

Several strands of work employ models for expressing legal requirements and assessing whether and to what extent these requirements are met by a given system. These strands include the large body of research concerned with the application of goal models to laws and regulations, e.g., [17,21], as well as a number of conceptual modeling techniques aimed at representing the semantics of legal texts, such as key legal abstrac-

tions and modalities, e.g., [43,36,2], and the structural representation of legal texts, e.g., [12,5,30].

As we discuss in more detail in Section 8, existing model-based approaches for compliance verification have one of the following limitations as far as the GDPR is concerned: they (1) have a different focus than the GDPR [26], (2) present guidelines only for the manual application of the GDPR [3], or (3) focus exclusively on specific GDPR use cases [8,27]. To the best of our knowledge, there are no proposals in the literature aimed at providing a holistic model-based representation of the GDPR. In order to address this gap, in a previous conference paper [42], we tackled the following three research questions (RQ):

- *RQ1: How can we develop a generic and adaptable model-based representation of the GDPR to support automated compliance checking?*
- *RQ2: How can we tailor the generic GDPR model according to the specific needs of a given context?*
- *RQ3: What are the challenges in modeling the GDPR?*

This submitted article is a major extension of our previous paper at MODEL 2019 [42]. In summary, the article enhances our earlier publication by providing: (1) the nine packages of the GDPR conceptual model developed in Enterprise Architect, (2) a table capturing how the classes in these packages are traceable to the GDPR (96 entries), (3) the complete glossary for our conceptual model (267 entries), (4) the plain-English description of 35 compliance rules derived from GDPR, (5) an encoding of said rules in OCL, and (6) a set of 20 variation points derived from the GDPR to specialize the generic model along with guidelines on how to apply them. None of these six complete artifacts were previously discussed in [42].

The complete material regarding points 1-6 above can be found in the publicly available Appendices A-C [40]. In particular, points 1-4 above are included in Appendix A, point 5 is presented in Appendix B, and point 6 is discussed in Appendix A (in plain-English) and Appendix C (in OCL with variability points).

Contributions. Our contributions are as follows: *(1)* We present the generic model of the GDPR composed of nine UML class diagrams and 35 OCL constraints. We use the term "generic" to imply that the model is based only on the content of the GDPR and is not encompassing any complementary information that may be necessary to contextualize the GDPR for use in a particular situation.
(2) The exact realization of the GDPR is subject to some variability depending on context. We present guidelines for tailoring the generic GDPR model into a specialized model that is suitable for application in a specific context. To this end, we describe 20 variation points that are considered acceptable by the GDPR and our strategy for handling these variations.
(3) We reflect on the lessons learned from encoding the GDPR into a model-based representation. Our lessons, which cover model validation, traceability and contextualization, provide a useful stepping stone for UML-based specification of other complex laws and regulations.
(4) We present the challenges we identified during our modeling endeavor alongside a number of future directions aimed at addressing these challenges.

Structure. Section 2 introduces basic concepts related to the GDPR. Section 3 provides an overview of our approach. Section 4 addresses our research questions. Section 5 and 6 present lessons learned and future directions, respectively. Section 7 discusses limitations and threats to validity. Section 8 compares with related work. Section 9 concludes the paper.

2 GDPR overview

The GDPR [15] is a complex piece of legislation comprised of 173 recitals, and 99 articles divided into 11 chapters. The GDPR applies primarily to businesses established in the EU. However, the regulation may also apply to businesses outside the EU, e.g., when these businesses offer goods or services to, or monitor individuals in the EU. If a business is subject to the GDPR, it has to identify itself as either a data controller or data processor. A controller determines the purpose and means of the processing, whereas a processor acts on the instructions of the controller. The responsibilities of a given business under the GDPR vary depending on whether it is a processor or a controller and depending on the kind of data processed.

Processors notably have to: (i) implement adequate technical and organizational measures to keep personal data safe and secure, and, in cases of data breaches, notify the controllers; (ii) appoint a statutory data protection officer and conduct a formal impact assessment for certain types of high-risk processing; (iii) keep records about their data processing; and (iv) comply to the GDPR restrictions when transferring personal data outside the EU.

In comparison to processors, controllers are subject to more GDPR obligations. In particular, in addition to having to meet the obligations mentioned above, controllers have to: (i) adhere to six core personal data processing principles, namely, fair and lawful processing, purpose limitation, data minimization, data accuracy, storage limitation, and data security; (ii) keep identifiable individuals informed about how their personal data

will be used; and (iii) preserve the individual rights envisaged by the GDPR, e.g., the right to be forgotten and the right to lodge a complaint.

3 Towards a Model-based Approach for Automated GDPR Compliance Checking

Our approach for enabling automated GDPR compliance checking has four steps, as depicted in Fig. 1.

Step 1 is a manual, one-off task aimed at building a generic model of the GDPR with the help of legal experts. More specifically, the goal of this step is to build, using UML class diagrams and OCL, a context-independent representation of the GDPR that does not take into account specific situations where EU member states' national laws, case law, or domain/organization decisions may affect the operationalization of the regulation. In this step, we develop, through a qualitative study, the following: (i) a generic model of the GDPR's main concepts and relationships, (ii) generic OCL constraints that verify GDPR compliance, (iii) a glossary to facilitate the understanding of the GDPR, and (iv) the variation points describing specific situations where the generic representation needs to be adapted to a given domain or organizational context.

In step 2, we tailor the generic model and OCL constraints of step 1 into a specialized model and a (specialized) set of OCL constraints. The goal of step 2 is to build an actionable basis for implementing the GDPR according to (i) the national laws of EU member states, (ii) GDPR case law, and (iii) other contextual information that may complement the GDPR. Step 2 yields two outputs that will later enable automated compliance checking in step 3. These outputs are: (i) a specialized model that represents the model tailored according to the application context, and (ii) a set of specialized OCL constraints which contain revised versions of the generic constraints developed in step 1 and potentially new constraints.

Step 3 is concerned with the generation of instances of the specialized model obtained from step 2. This is done via a model-editing tool that allows legal experts to create representations of legal documents (e.g., a privacy policy statement) or GDPR-related information extracted by databases in the form of an instance of the specialized model. This is an intermediate step to create representations of legal and technical documents in the form of an instance of the specialized model. Stated otherwise, step 3 generates a model instance providing a structured representation of the legal and technical documents that have a bearing on GDPR compliance. As an example, consider the model representation for

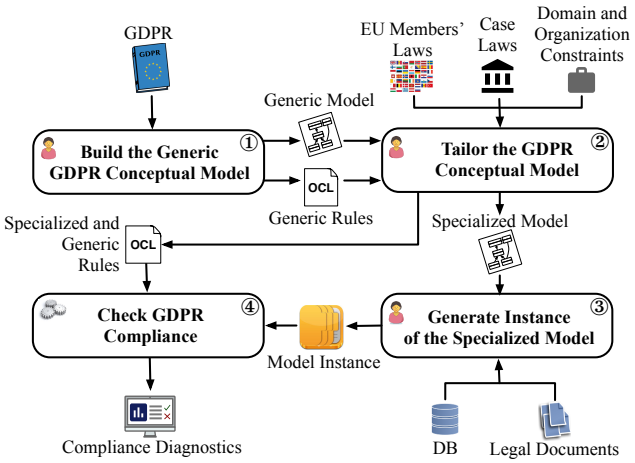

Fig. 1 Approach for Automated GDPR Compliance Checking

privacy policy statements we introduced in another recent piece of work [39]. This latter model is an instance of the specialized model obtained in step 2.

Finally, in step 4, the model instance generated from step 3 is checked against the specialized OCL constraints obtained from step 2. For example, in our recent work [39], first we selected the OCL constraints created for privacy policies. We then specialized those OCL constraints for the Funds domain (step 3). Finally, the specialized constraints were used to inspire the creation of completeness criteria to check if a given privacy policy is compliant with GDPR. The compliance diagnostics resulting from the constraint checking process are then delivered to end-users, typically legal experts, in a user-friendly manner.

In this paper, we describe our experience conducting steps 1 and 2. Steps 3 and 4 are discussed in another recent publication [39]. Steps 1 and 2, along with their inputs and outputs, are discussed in detail in Sections 4.1 and 4.2, and in the Appendices A-B [40].

4 Modeling the GDPR

4.1 Building a Generic Model for the GDPR (RQ1)

In the first step of our approach (Fig. 1), we build a generic model representing the GDPR without accounting for the specificities of the application domain. This modeling activity addresses RQ1 and yields: (1) a UML Class Model (CM) that captures the GDPR's key concepts and their relationships (see Section 1 of Appendix A); (2) a set of 35 OCL constraints over the CM reflecting the GDPR's obligations (see Section 3 of Appendix A for the plain-English version, and Appendix B for their OCL version). Given a specific context, the applicable constraints need to be completed so that one

can evaluate them in an automated and precise manner; (3) a glossary of 267 terms to understand the GDPR model (see Section 2 of Appendix A); and (4) a table that summarizes all variation points extracted from the GDPR (see Section 4 of Appendix A). The output table mentioned above aims to facilitate the work of analysts in the subsequent tailoring step (Section 4.2). Below, we explain the methodology we employed to create these outputs. We then illustrate the outputs using concrete examples.

Modeling methodology. This modeling activity was performed in an iterative and incremental manner as shown in Fig. 2. Each iteration was interleaved with a thorough validation session with experts from the legal domain, noticing that those experts were already trained to understand the CM notation. The team was composed of three legal experts: (a) one was a senior lawyer with more than 30 years of experience in European and international law; (b) one was mid-career lawyer with more than 10 (but less than 20) years of experience in law in the financial domain; and (c) one was an IT professional with more than 10 years of experience in the legal domain. Building the generic model for the GDPR took four iterations with each iteration requiring on average two weeks. The first three authors of this paper built the generic model with the help of the legal experts. To mitigate biases, the modeling activity was systematically performed by a pair of researchers (the first and second, or the second and third, or the first and the third authors of this article). Afterward, the third researcher (the one that was not involved in a the activity) reviewed and challenged some of the results of a given modeling activity. Finally, the fourth and fifth authors participated in meetings and provided feedback throughout the modeling endeavour. This activity was performed over 6 months in an iterative and incremental manner with face-to-face, bi-weekly sessions with the team of experts, each of these sessions lasting between two to three hours. Each session was attended by at least three out of the five authors and at least two out of the three experts. Each session started with the authors in attendance presenting to the experts, using slides or printed documentation, the new parts of the GDPR that had been modelled in the form of diagrams and/or natural language descriptions of OCL constraints. Subsequently, the experts were invited to provide feedback. The discussions continued until the experts in attendance agreed that the models and constraints correctly reflected their interpretation of GDPR. In addition to face-to-face meetings, we had several off-line validation sessions with legal experts, which approximately took an additional 20 hours. This activity was concluded when the experts did not have any additional suggestions for improving the clarity, completeness, or correctness of the model.

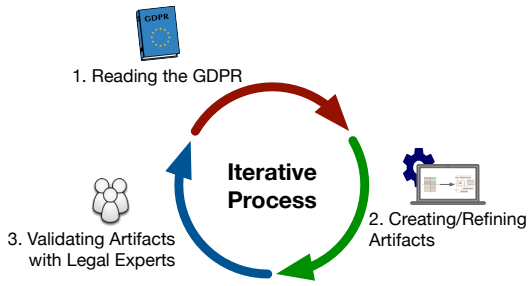

Fig. 2 Iterative Process.

During the first iteration, we read the GDPR in its entirety and tried to extract important definitions, concepts, rules and possible variations from it. Figure 3 illustrates the information extracted from Art. 8 – the article regulating how a child data subject can provide consent for processing her personal data in the context of information society services. In particular, eleven concepts (shaded gray), two rules, and one variation point were extracted from the excerpt of Art. 8 in Figure 3. Recent work uses natural language processing techniques to extract such legal information in an automated manner [33]. Nevertheless, we opted for a manual strategy to avoid overlooking any important information while deepening our understanding of the GDPR. Among other reasons, a manual strategy was essential for enabling the identification of GDPR rules in a fully precise manner. For example, we have mapped the rules in Art. 8 to their corresponding OCL constraints as we illustrate later.

Based on the extracted information, and using our understanding and interpretation, we created the modeling artifacts listed earlier. Next, these artifacts were presented to legal experts for feedback. In addition to pointing out issues and omissions, our collaborating legal experts were encouraged to bring to our attention any GDPR article that they suspected might have been misinterpreted, i.e., incorrectly modeled. By doing so, we boosted subsequent iterations since we no longer needed to analyze the entire GDPR again.

In practice, we observed that the corrections suggested by the legal experts were, by and large, based on conventions or articles that were not part of the GDPR itself, e.g., articles from the Article 29 Working Party (WP)[1]. For example, a data controller might need to si-

[1] Art. 29 WP is the independent European working party that dealt with issues relating to the protection of privacy and personal data until 25 May 2018 (date at which the GDPR took effect). All archives from Art. 29 WP are available at: https://ec.europa.eu/newsroom/article29/news-overview.cfm. Art. WP 29 has been replaced by the European Data Protection Board; see https://edpb.europa.eu

multaneously communicate with many supervisory authorities; such authorities are established by individual European member states to supervise compliance with the GDPR. In such a case, the controller has to designate a unique *lead* supervisory authority (Art. 56). Subsequently, the controller should only communicate with the lead supervisory authority, which in return, will coordinate any investigation or administrative task with the other concerned authorities. Although not explicitly stated in the GDPR, the choice of the lead supervisory authority is not arbitrary. The lead supervisory authority should be selected based on predefined rules that account, among other things, for the location of the main establishment of the controller and where the actual data processing is taking place (Working package 244 of the WP).

In the next modeling iteration, we re-read the GDPR parts and other annex documents that were noted by the legal experts in the previous iteration. Then, we refined the outputs according to expert feedback, and so on. Once the conceptual model started to stabilize, we put together a general report including all the resulting outputs for off-line validation. The modeling step terminated when the general report, containing all the model artifacts, was approved by the legal experts.

Illustration of the modeling artifacts. Fig. 4 depicts the package view of the CM. To keep the CM manageable and easy to grasp as it grows in size, we spread the CM classes over nine packages as follows, noting the package names are self-explanatory.

This work covers seven out of the eleven chapters presented in the GDPR. Packages *GDPR Principles*, *Data Subject Rights*, and *Data Transfer* respectively cover chapters 2, 3, and 5 of the GDPR. Concepts from chapters 1, 4, 8 and 9 were spread over the remaining packages based on their meanings and roles. For example, concepts from chapter 4, which is the longest chapter and where most GDPR compliance requirements are defined, are grouped in packages *Data Processing*, *Compliance Evidence*, and *Actors*. Chapters 6, 7, 10, and 11 have little to no impact on compliance checking, and subsequently were excluded after the first modeling iteration. For example, chapter 6 regulates the internal functioning and composition of the public data supervisory authorities. The nine CM packages, their traceability with GDPR, and their description are presented in Appendix A. In Fig. 5, as an example, we show an excerpt of the *Data Processing* package that covers most concepts extracted from Art. 8 in Fig. 3.

Intuitively, the CM in Fig. 5 presents the information that has to be collected when the lawfulness of data processing is based on consent. In the CM, only data processing manipulating some personal data should be considered. Other kinds of processing are out of scope. The purposes for each processing have to be explicitly defined (see *realizes* association between *Data Processing* and *Purpose*), noting that several instances of processing can share one or more purposes. A well-designed consent form should, among other things, remind data subjects of all their applicable GDPR rights. Consent is given by data subjects, or their responsible parent in case of a child data subject, for one or more predefined processing purposes (see *given for* association between *Consent* and *Purpose* and *gives* association between *Data Subject* and *Consent*). This is only possible when the treated personal data is sufficient for the precise identification of data subjects (see *identifies* association between *Personal Data* and *Data Subject*).

The CM comes with 35 constraints, as presented in plain English in Section 3 of Appendix A, and in OCL in Appendix B. OCL constraints are expressed as invariants denoting logical conditions that must always hold over all instances of a given class. Listing 1 presents three OCL constraints related to excerpt of the *Data Processing* package in Fig. 5.

For example, the invariant named C5 checks that when lawfulness is based on consent (L. 2), the consent for child data subjects has to be provided by their legal responsible parent (L. 3-14). This constraint involves no variability and does not require any additional tailoring in the subsequent step.

```
1  context Data_Processing inv C5:
2  self.isLawfulnessOnlyByConsent() implies
3  let identifiableSubjects : Set(Data_Subject) =
       self.personal_data.data_subject->flatten()->
       asSet() in self.purposes->forAll(p:Purpose|
4      identifiableSubjects->forAll(ds: Data_Subject|
5          let eligibleToGiveConsent:Natural_Person =
6          if ds.oclIsTypeOf(Child_Data_Subject)
7              then ds.getResponsibleParent()
8              else ds endif in
9          p.getConsents()->forAll(c:Consent|
10             c.provider=eligibleToGiveConsent
11             and c.target=ds
12         )
13     )
14 )
15 context Data_Subject inv V1:
16     let minDSAge: Integer = Variability.
       V_getMinimumAgeForDS(self) in
17     if (self.oclIsTypeOf(Child_Data_Subject))
18         then self.getAge() < minDSAge
19         else self.getAge() >= minDSAge endif
20 context Natural_Person inv V2:
21     self.children - > forAll(c: Child_Data_Subject
         | self.V_checkParentDocuments(c))
```

Listing 1 Examples of OCL Constraints

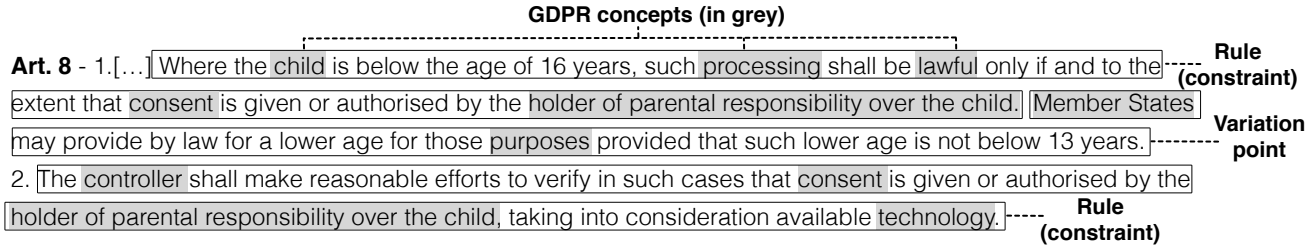

Fig. 3 Example of Information Extracted from (Excerpt of) Article 8 of the GDPR

Constraints involving variability are distinguishable by their name, which includes the "V" prefix, e.g., V1. We handle variability in OCL constraints using partially specified operations that need to be later updated or redefined based on the context at hand. For example, the second constraint (L. 15-19) states that the age of data subjects should be greater than a certain dynamic threshold. However, when the context is known, the operation V_getMinimumAgeForDS should dynamically identify the value of the threshold based on the country of residence of the data subject and the locations of the involved data processing, controllers, and processors. Finally, V2 (L. 20-21), checks that a given person is indeed the holder of parental responsibility over a given child data subject. We further discuss variability in Section 4.2.

To ease the understanding of the modeling artifacts for legal experts, we rely on a glossary of important terms. The glossary has 267 entries for the CM (see Section 2 of Appendix A). In addition, we include intuitive descriptions for each OCL constraint (see Section 3 of Appendix A). Table 1 presents an excerpt of the glossary that supports the CM package in Fig. 5. The plain English description of the OCL constraints in Listing 1 and their traceability to the GDPR can be found in Section 3 of Appendix A. The first column of Table 1 points to the modeled concept (i.e., term in the table) such as the classes. The second column presents an intuitive natural-language description of the element in the first column. For example, the class Data Processing is described in the second row of Table 1. In addition to the description, we present in Section 1 of Appendix A the GDPR source articles of the elements of the CM packages showed in Figure 4. Here, traceability is meant to help legal experts during the validation sessions. In particular, it makes it easier to spot whether we have missed some important articles that might further consolidate the definition of a given concept.

The last modeling artifact is a table including all possible variation points extracted from the GDPR. We defer the discussion and presentation of this table to Section 4.2.

4.2 Specializing the Generic Model (RQ2)

In the second step of our approach (Fig. 1), analysts tailor the generic modeling artifacts to account for the specific context and activities of the organizations seeking compliance. This step addresses RQ2.

Generally speaking, analysts have to resolve all the variations that are relevant to the context at hand. This might introduce new constraints coming from the specific law of a European Member State (EMS), GDPR case laws, and other contextual information that may complement the GDPR. In this paper, we focus on addressing the variability that may come from the specific law of a EMS as expressed in the GDPR.

The output of this step is a specialized and augmented version of the modeling artifacts created in the first step of our approach (Section 4.1). As mentioned in Section 3, the variability in the GDPR comes from the fact that the interpretation or the enforcement of some provisions may be affected by additional acts and laws from the EMS.

Table 2 presents an excerpt of the variability table that contains five variation points (two of them, namely V1 and V2, used during the example showed in the first step of our approach). The complete set of 20 variation points is presented in Section 4 of Appendix A. These variation points were identified from the GDPR by the authors of this paper. To capture the variation points, we focused on the GDPR provisions that enable the EMS to adapt GDPR provisions to its specific laws. Those variation points were completed with the help of legal experts. The experts had complete purview of the way GDPR had been negotiated, including the flexible points that had been provisioned to accommodate all the EMS as well as the current case law (circa 2019) related to the GDPR and how GDPR was being interpreted in different contexts and EMS. This table will

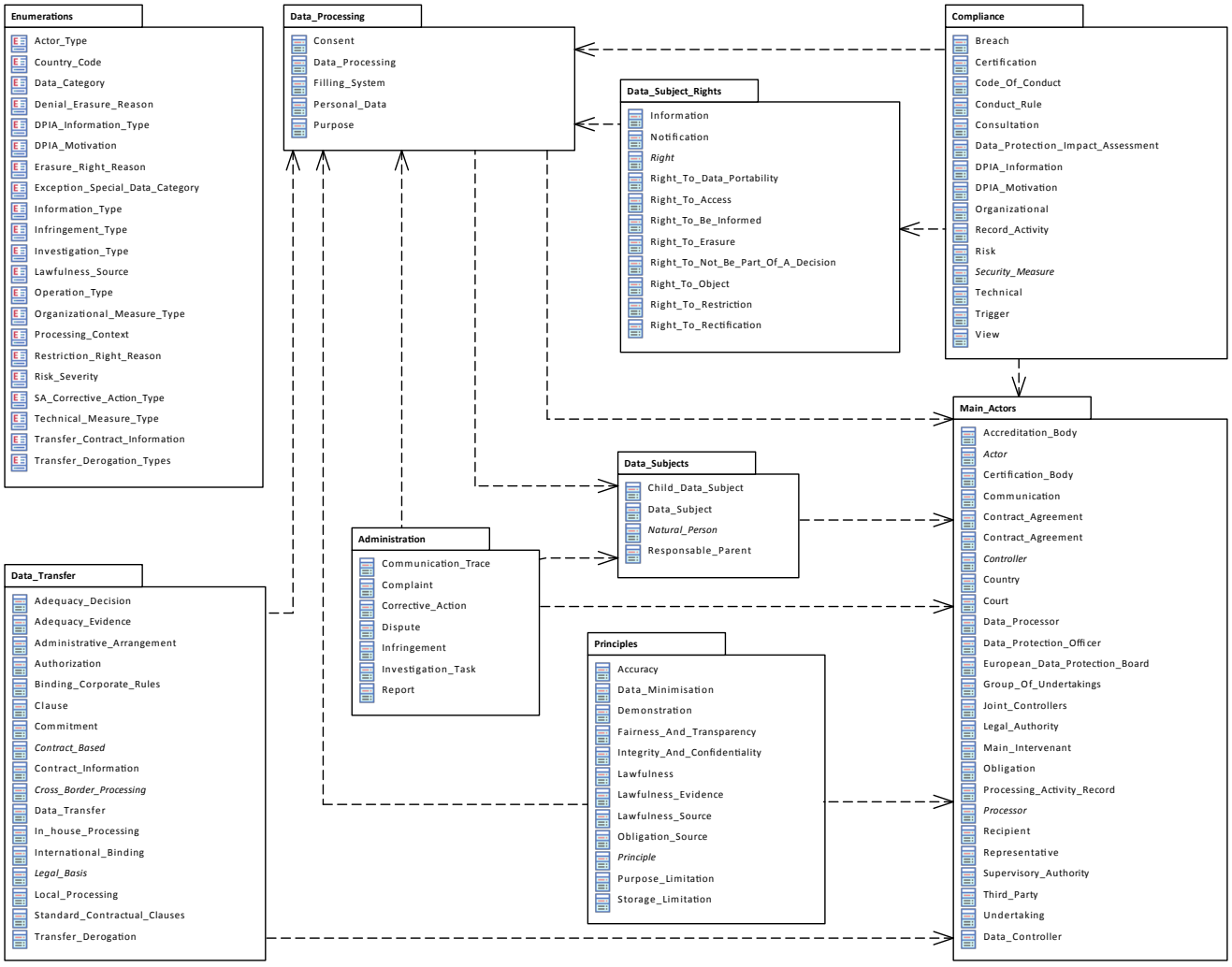

Fig. 4 Package Representation of the CM

guide analysts in better understanding when and how they should resolve a given variability.

The first column of the Table 2 represents the identifier of the variation point. This identifier will be later used to point to the OCL constraint or model adaptation that is needed to solve a given variation (see Appendix C). For example, the variation point V1 is addressed by the OCL constraint V1 shown in Fig. 3. The second column of Table 2 traces the variability to the GDPR. The third column provides an intuitive textual description of the variation. The third column indicates also the actor that should be consulted for resolving the variation, e.g., the EMS. Note that the description also covers when the underlying actor is likely to influence the interpretation and enforcement of the original GDPR rules. For example, in V10, the EMS law may, for important reasons related to public interest, expressly set limits to the transfer of specific categories of personal data to a third country or an international organisation. At the time this article was written, data can be transferred within the same international organization to Switzerland without additional obligations. However, unconditional data transfer to third countries is limited, for example, transfer to Canada is only limited to commercial organizations under Canadian's PIPEDA law (Personal Information Protection and Electronic Documents Act). Other sectors and domains involve additional obligations that need to be fulfilled such as the approval of the lead supervisory authority. The fourth column of Table 2 provides the context required by analysts to understand how they should resolve the variation.

The strategy we employ for resolving most of the variation points is "clone and own" [10], where the

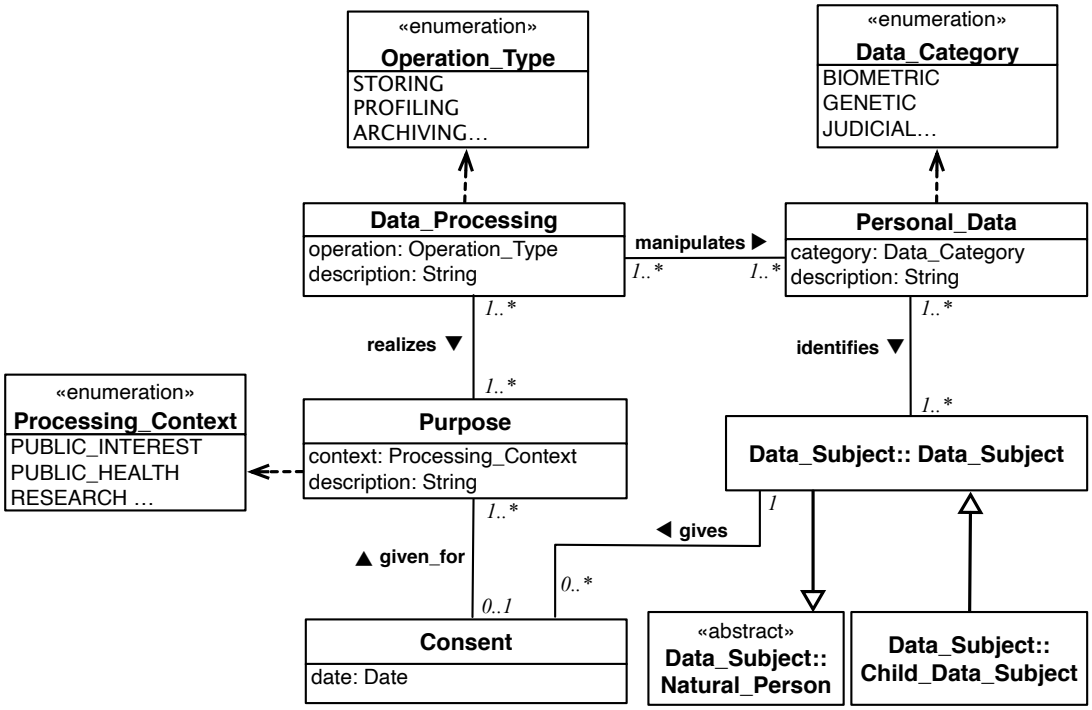

Fig. 5 Excerpt of the *Data Processing* Package

Table 1 Glossary Excerpt

Term	Description
Personal Data	Personal data means any information relating to an identified or identifiable natural person.
Data Processing	Is any operation performed on personal data, whether or not by automated means, including collection, recording, organization, structuring, storage, etc.
Data Subject	A natural person whose personal data is processed by a controller or processor.
Consent	It means any freely given, specific, informed and unambiguous indication of the data subject's wishes by which he or she, by a statement or by a clear affirmative action, signifies agreement to the processing of personal data relating to him or her.
Purpose	The purpose of data processing is said to be lawful if its legal basis matches one of the possible circumstances under which GDPR permits the processing of personal data. Example of valid legal basis for data processing are consent and when processing is necessary to perform or prepare for a contract with the data subject.

generic artifacts are specialized for the organization and system(s) at hand. Examples of changes to the artifacts include updating the cloned CM, glossary, and the OCL constraints. Further, analysts can add new OCL constraints, and drop or override existing ones. The only artifact from the first modeling step of our approach that remains unchanged is the variability table (Table 2). This is because the variability table incorporates all possible variations with regard to the GDPR and is used as a checklist for guiding the analysis during the tailoring step. Concretely, analysts skim through the variability table and resolve the variations that apply to the underlying context. An important challenge here is keeping track of the changes made for specializing the modeling artifacts. To do so, analysts have to record the actions they have taken to tailor the generic modeling artifacts. To illustrate, let us suppose that an organization X is an international commerce company located in Europe and Canada processing sensitive personal data.

Table 3 presents an example of how the variability in Table 2 would be handled for X. The first column of Table 3 references a particular variation ID listed in Table 2, whereas the second column of Table 3 lists the cloned generic artifacts that were impacted during the resolution of the variation. The final column of Table 3 describes how the artifacts in the second column were updated based on the specific context of X. X must account only for V3 and V4 in Table 2. V1 and V2 in Table 2 do not apply to X since X only trades with subjects aged over 18 years old (clearly stated in X's privacy policy and website). V10 does not apply to X since X has an adequacy decision. X is also requested to conduct a DPIA (Data Privacy Impact Assessment) to be able to perform cross-border data transfer to Canada with an adequacy decision. The specific adequacy deci-

Table 2 Excerpt of the Variability Table

ID	Source	Description	How to resolve
V1	Article 8	EMS law may provide for a lower age (. . .) provided that such lower age is not below 13 years.	Add the V1 OCL constraint to the generic model and implement V_getMinimumAgeForDS based on the EMS laws.
V2	Article 8	Checks that a given person is indeed the holder of parental responsibility over a given data subject according to the EMS law.	Add the V2 OCL constraint to the generic model and implement VcheckParentDocuments based on the EMS laws.
V3	Article 9	The processing of sensitive personal data is prohibited unless the data subject has given explicit consent (. . .), except where (. . .) EMS law provide that the prohibition (. . .) may not be lifted by the DS	Implement V_prohibitionCanBeLiftedByConsent of the constraint C6 based on the EMS laws.
V4	Article 9	EMS may maintain or introduce further conditions,including limitations, with regard to the processing of genetic data, biometric data or data concerning health	Add the V4 OCL constraint to the generic model and implement VcheckParentDocuments based on the EMS laws.
...	...	...	...
V10	Article 49	In the absence of an adequacy decision, (. . .) EMS law may, for important reasons of public interest, expressly set limits to the transfer of specific categories of personal data to a third country or an international organisation.	Add the V10 OCL constraint to the generic model and implement V_verifyTransferLimits based on the EMS laws.

Table 3 Variability Resolution Table

Ref.	Artifact	Summary of actions
V1	-	[Not applicable]
V2	-	[Not applicable]
V3	Specialized Model	The specialized model includes an update version of the constraint C6.
V3	OCL constraints	Implement V_prohibitionCanBeLiftedByConsent of the constraint C6 based on the EMS laws.
V3	Glossary	Add the terminology of the implementation of V_prohibitionCanBeLiftedByConsent to the glossary.
V4	Specialized Model	The specialized model includes the new constraint V4.
V4	OCL constraints	Add the V4 OCL constraint to the generic model and implement V_verifyFurtherConditionsAndLimit based on the EMS laws.
V4	Glossary	Add the terminology of the new constraint V4 and the implementation of V_verifyFurtherConditionsAndLimit to the glossary.
...	...	...
V10	-	[Not applicable]

sion for Canada is called PIPEDA. Other generic constraints (i.e., C27, C28, C32, C33 among other rules) handle the DPIA and the adequacy decisions.

As shown in Table 3, only the variation points V3 and V4 are relevant for X. Specifically, the cloned and possibly specialized model, OCL constraints, and glossary were altered as described in the last column of the table. For example, to address V3, the OCL constraint C6 that checks that the processing of special data categories are prohibited unless the processing is based on consent needs to be updated by implementing V_prohibitionCanBeLiftedByConsent if the the prohibition can be lifted by consent based on the EMS laws (third to fifth row of Table 3). Instead, to address V4, the new OCL constraint V4 needs to be added to encode the variation point. In addition, the analyst has to implement V_verifyFurtherConditionsAndLimit in V4 as it is imposed by one of the EMS laws that are relevant to X (sixth to eight row of Table 3).
The analyst that is tailoring the specialized model should take into account two main points. First, regardless of the changes made, the OCL constraints should remain correct with respect to the cloned and possibly specialized CM. For example, if the analyst decides to drop the class *Consent* and its associations, then all impacted constraints have to be either corrected or dropped. Second, the analyst might unintentionally introduce inconsistencies in the set of OCL constraints, e.g., two contradicting constraints. To avoid this, one can employ existing constraint solvers, e.g., UML2CSP [7], Alloy [7] or PLEDGE [35], to spot UNSAT sets of constraints, and consider if any of the UML consistency rules discussed in the literature could apply to the application context [41].

4.3 Challenges Encountered during the Modeling (RQ3)

In this section, we address RQ3 by listing the main challenges encountered when modeling the GDPR. Later, in Section 6, we present our vision for how we intend to address these challenges.

Specification of Compliance Rules (Challenge 1): In this paper, we first use OCL constraints to embed compliance rules in the generic model. We then adapt and expand these constraints to create a specialized model.

We have already taken care of defining OCL constraints over the generic model; no additional effort is thus foreseen for this task. Nevertheless, additional effort, including by legal experts, will be required for defining OCL constraints over the specialized model, noting that these constraints necessarily refer to legal material (e.g., EMS laws, GDPR case law, and domain adaptations) that is more complex and fragmented than the GDPR. Due to the scarce familiarity of legal experts with OCL, the creation of the latter group of constraints may be difficult and time-consuming.

Rationale for Model Specialization (Challenge 2): Although we keep track of all the actions performed during the tailoring step, we do not systematically express the rationale behind the actions; in other words, we do not document why analysts made the decisions they did [31]. In the context of our work, the rationale needs to cover the problems the analysts encountered, the options they investigated, the GDPR provisions they examined to evaluate the options, and, most importantly, the arguments that led them to make certain decisions.

Generation of the Instance Model (Challenge 3): The process of generating an instance of a specialized model is currently dealt with manually (recall step 3 in Fig. 1). This would mean that a legal expert would have to create, by using a model editor, the instance. This manual process is time-consuming and tedious.

5 Lessons Learned

In this section, we discuss the lessons we learned from modeling the GDPR.

Streamline the validation process. We observed that modeling the GDPR, whether at a generic or specialized level, necessitates substantial legal knowledge and expertise that may go beyond the GDPR itself, e.g., knowledge of the Article 29 Working Party. Thus, putting in place an effective and efficient validation process with legal experts was paramount to ensure that the produced artifacts were as complete and precise as possible. To achieve this goal, we had to shield the legal experts from the complexity arising from the CM and its underlying OCL constraints.

As discussed in Section 4.1, legal experts were able to grasp the CM with relative ease. This was in large part thanks to the intuitiveness of UML class diagrams and the fact that non-software experts can be quickly trained to obtain a working understanding of the notation for validation purposes. In general, we observed from experience that class diagrams can be taught to non-IT experts with relative ease [34]. In contrast, OCL, which we use to formally express the GDPR rules, was challenging and intimidating to legal experts, despite our attempts to explain the meaning of the constraints. Similar communication barriers were observed when we attempted to replace OCL with other logical notations, e.g., standard first-order logic. In general, we believe such barriers are to be expected when formal logic is used directly with professionals who do not have adequate mathematical background. We mitigated this issue by describing each OCL constraint via an intuitive but precise textual description in natural language (see Section 3 of Appendix A). Nevertheless, the plain-English explanation of the OCL constraints per se was still not enough to ensure reliable validation of the OCL constraints. In particular, the same rule can be often expressed over smaller and modular sub-constraints. For example, the excerpt of the article of Fig. 3 was encoded over three constraints, namely V5, V1 and V2 in Listing 1. The former constraint encodes the common part of the rule, whereas the latter two (variation points constraints) cover the variable part. With the rules getting fragmented, legal experts experienced difficulties because they could no longer relate to the original rule. One way to remedy this problem is by forcing one-to-one mappings, where any GDPR rule is expressed using a unique OCL constraint. However, such a solution will further complicate the tailoring step, since variant requirements will have to be mixed with the fixed ones. This prompted us to create (1) a plain-English constraints table (see Section 3 of Appendix A) which traces the GDPR rules to their corresponding constraints, and (2) the variability table (see Section 4 of Appendix A) that contains the variation points expressed in plain-English traced to the GDPR. Both previous tables facilitated the validation of the OCL constraints by legal experts.

The variability table (e.g., see Section 4 of Appendix A) was enough to enable the legal expert to verify that the list of extracted variation points was complete and precise. A simple but effective solution was to support several views for the same CM, where the level of detail to display is configured according to needs. Although the validation of the CM was conducted package by package, legal experts still found the models to be overwhelming in terms of their information content. To this end, we found out that, in many situations, hiding class operations, attribute types, and stereotypes would be helpful. Further, to ensure that enough time was given for validation, we alternated online and off-line validation as discussed in the modeling methodology of Section 4.1.

Maintain traceability. Another observation from our GDPR modeling experience is that both analysts and

legal experts often needed to consult specific articles to refresh their memory. Being able to do so effectively required all our modeling artifacts to be traceable to their corresponding GDPR provisions. Examples of traceability links can be seen in the second column of the tables in Section 1 of Appendix A, the plain-English constraints table (Section 3 of Appendix A) and the variability table (Section 4 of Appendix A). Although not shown in Fig. 5, classes too are traceable to the specific GDPR provisions pertaining to them at the level of the CM. For example, the class *Purpose* in Fig. 5 is mapped to Arts. 5, 13, 14 and 15. These links made it easy to go back and forth between the modeling artifacts and the GDPR. We anticipate the links to be useful for other purposes as well, e.g., performing impact analysis when the GDPR, or the EMS laws change. The only classes that were not mapped to the GDPR are those we have created to better structure the CM, e.g., the *Processing Activity Record* class.

A further final observation about traceability concerns the importance of maintaining consistent relationships between the different modeling artifacts. In practice, one often needs to quickly navigate from one artifact to another, in particular during the tailoring step. For example, when resolving a given variability, it is often useful to view the list of rules whose fulfillment is likely to be impacted by the EMS laws. Similarly, analysts need to navigate to the underlying OCL constraints that need to be updated. Examples of such links can be found in the third column of plain-English constraints tables (Section 3 of Appendix A) and the second column of variability tables (Section 4 of Appendix A). We received positive feedback from the legal experts about having such navigable artifacts. In particular, legal experts appreciated the intuitive and GDPR-traceable approach of our model artifacts that allows them to maintain the integrity of the artifacts by reducing arduous tasks such as going back to read a specific article of the GDPR related to a class (without traceability).

Make the tailoring step as systematic as possible. During the tailoring step, we observed that even experienced analysts could encounter difficulties in resolving the variation points. The root cause of this was the large number and size of the modeling artifacts. This prompted us to develop simple guidelines to systematize and better organize the tailoring step. First, analysts have to go through the variation points and tick those that are relevant to their working context. Then, analysts can focus only on the relevant variation points and apply our recommendations on how to resolve them. This was facilitated by the "How to resolve" column of the variability table (see Section 4 in Appendix A).

For example, when V1 in Table 2 is relevant to the context, analysts will get to know that they have to update V_getMinimumAgeForDS to account for the minimum age of children as regulated by the relevant EMS laws. However, we do not recommend a sequential resolution of variation points, e.g., first resolving V1, then V2, then V3, and so on. In particular, analysts should postpone completing the specification of the OCL constraints until all the variability for the CM has been handled. This is because some changes in the CM might break other constraints for which variability was previously resolved. To help analysts follow these recommendations, we proposed to keep track of all the tailoring actions in the resolution table (see Table 3). This facilitates resolving the variabilities in an incremental and non-sequential manner. In line with the above, a recent work from Hajri et al. proposes a tool-supported approach that guides analysts in configuring product specific models from product line models [20] [19]. In the future, we envisage to operationalize our tailoring recommendations by customizing Hajri et al.'s work.

Finally, we found the resolution table to be very useful when we had to deal with several similar contexts. In such cases, we started the tailoring from specialized modeling artifacts produced for other similar contexts, rather than from the generic artifacts. This, in our experience, helps to expedite the tailoring step.

6 Future Directions

In this section, we describe the most important future directions that, we believe, are necessary for addressing the challenges identified in Section 4.3.

Domain-Specific Rule Language (Challenge 1). Using OCL constraints is key to achieving automation in checking GDPR compliance. In our approach, this is done via the specialized set of OCL constraints that encode the rules applying to a given context. Nevertheless, some of the specialized constraints, in particular, the new ones originating from the EMS laws, have to be validated by legal experts. As discussed in Section 5, OCL impedes understandability by legal experts. To tackle this limitation and improve the tailoring of the generic model (Step 2 in Fig. 1), it would be advantageous to develop a Domain-Specific Rule Language (DSRL). The DSRL should, on the one hand, be expressive enough to be useful for the precise specification of GDPR compliance checking rules, and on the other hand, understandable enough to be readily used by legal experts. For example, the OCL rule presented

in Listing 1, would be hardly understood by most legal experts. To ease understandability, restricted natural language (NL) could be used as the basis for the DSRL. While basing the DSRL on NL increases usability, there is still the risk that legal experts may find it difficult to articulate their rules in a proposed language. To mitigate this issue, one needs to closely interact with legal experts during the DSRL design, and iteratively validate the language constructs with them. In addition, providing training material for the DSRL would be essential to make the language more accessible to non-software experts. Finally, to support automated compliance checking, the rules specified in the DSRL should be automatically translatable into OCL so that the rules can be checked directly over instantiations of a specialized GDPR model.

Goal Models (Challenge 2). Using goal models can help to deal with capturing and reasoning about the rationale for model specialization. Each goal is a prescriptive statement of intent that a system should satisfy [22]. Here, the term "system" refers to a combination of IT applications, organizations, work-flows and people that together perform certain functions. A goal model is characterized by a collection of goals, the relationships (e.g., hierarchical decomposition) between the goals, and the obstacles that could hinder the satisfaction of the goals. Goal models provide a flexible instrument for arguing about model specialization. A key task related to a goal-oriented analysis of the GDPR would be to decide how the application context discussed in Section 2 (Step 2 in Fig. 1) should be decomposed and analyzed in order to tailor the specialized model. This decomposition necessarily involves breaking down the GDPR's core tenets (e.g., data minimization) into more tangible sub-goals. Additionally, one may need to decompose the goals of a given system (e.g., a specific organization), and examine how the system goals map onto the goals stipulated by the GDPR. A main criterion to fulfill regarding goal decomposition would be to ensure that the decomposition process makes progress towards a set of concrete claims for which meaningful evidence about satisfaction (in term of model specialization) could be collected. Meeting this criterion necessitates that the developed goal models should provide a blueprint for the justification that is needed in order to argue about the adequacy and effectiveness of a proposed model specialization.

AI-enabled Automation Support (Challenge 3). Legal documents typically come in the form of NL descriptions. Mining these descriptions to identify the appropriate metadata to build the instance model is a prerequisite for automated compliance checking. Metadata items relevant to GDPR are numerous. Examples of such metadata include: "purpose" to mark the purposes of the processing for which personal data is being collected, "basis" to mark the legal basis for the processing of personal data, and "right to access" to mark the clause(s) giving an individual the right to request from the controller access to their personal data. These metadata items have to be identified in legal and technical documents such as privacy policies, consent statements, records of processing activities and exemptions, and data protection impact assessments. Natural Language Processing (NLP) [23] and Machine Learning (ML) [1] provide a useful technical platform for metadata extraction [33]. The metadata identified will be the basis for the model-based representation of the legal and technical documents to be checked. In other words, an automatic instantiation process will convert the metadata extracted with NLP and ML for a given document into a model-based representation, i.e., the instance of a specialized model. The elements of this instance model will be both fully traceable to the content of the source document as well as unambiguously mappable onto the underlying generic and specialized models. The initial results tackling this future direction are presented elsewhere in recently published work [39].

7 Limitations and Threats to Validity

In this paper, we draw on MDE for building a machine-analyzable representation of the GDPR as a first step towards the development of future automated methods for assessing GDPR compliance. In the next two subsections, we present the limitations and the threats to validity of our work.

7.1 Limitations

The limitations of our work include:

Generalizability. Our approach was built considering only the provisions of GDPR and cannot necessarily be applied to other privacy laws. Nevertheless, our qualitative methodology to build and validate the GDPR model can be reused to create other models that represent different laws. For example, assuming that the appropriate legal experts are available, a similar approach for the Personal Information Protection and Electronic Documents Act (PIPEDA), which is the Canadian law that regulates the collection, use and disclosure of personal information, could be developed by following the methodology presented in this paper.

Resiliency. If there are drastic changes to the GDPR over time, for example, the introduction of new concepts and provisions, we do not believe that either we, or

anyone else, would be able to develop a future-proof model. Evolution and alignment of our model may thus be necessary if GDPR evolves.

Extendability. We anticipate that future automated solutions for checking compliance of GDPR-related documents and systems will need to consider details that are not considered in our work at the moment. For example, consider the following details that were included in another recent work [39] where we automated the activity of checking GDPR compliance of privacy policies: (a) the concept of adequacy decisions between the EU and a territory (e.g., Andorra, the Bailiwick of Jersey, etc.), specific sectors (e.g., the commercial organizations from Canada, Argentina, etc.), and a country (e.g., Japan and New Zealand), (b) the legal basis contract and more specifically whether the provision of personal data for this legal basis is a statutory requirement, a contractual requirement, or a requirement necessary to enter into a contract, (c) the appropriate safeguards, i.e., binding corporate rules or EU model clauses that would allow the controller to share the collected personal data to recipients outside Europe, and (d) the specific derogation in terms of unambiguous consent to allow the controller to share the collected personal data outside Europe. In addition, our approach does not consider the notion of a system. For example, according to the security principle of the GDPR (see Article 5.1(f)), the data controller's system shall ensure appropriate personal data security, including protection against unauthorized or unlawful processing and accidental loss, destruction, or damage, using appropriate technical or organizational measures. Such a system should also implement appropriate technical and organizational measures to ensure a security level commensurate with the risk. Operationalizing our GDPR models in the context of secure systems and software development is not addressed in our current modeling endeavor and is left for future work. Our GDPR models will need to be integrated with actual system models and their requirements (e.g., related to security). Such integration is important for demonstrating traceability and completeness and verifying compliance.

7.2 Threats to Validity

Below, we discuss threats to the validity of our approach and what we did to mitigate these threats.

Internal Validity. A potential threat to internal validity is that the authors of this paper interpreted the text of GDPR provisions in order to create the generic model presented in Fig. 4. To minimize the threat posed by such subjective interpretation, this phase was done in close collaboration with independent legal experts (the Linklaters legal experts specialized in GDPR). While we cannot rule out subjectivity, we provide our interpretation in a precise and explicit form. In addition, our model is publicly available and thus open to scrutiny.

External Validity. Our qualitative study leading to the creation of our generic GDPR model was enhanced by feedback from legal experts who had familiarity with data protection in a variety of domains. This provides some degree of confidence about our results being generalizable. That said, future studies that instantiate our model in different legal domains will be essential for determining the completeness and general applicability of the model.

8 Related Work

In this section, we distinguish two categories of studies related to the work presented in this paper: proposals that report on (1) Modeling the GDPR, and (2) Checking Compliance. Table 4 provides the comparison of the 11 research papers analyzed in this section. The first column *"Reference"* of the table provides a reference to each study. The second column *"Year"* indicates the year when the study was published. The third column *"GDPR Coverage"* shows the degree of coverage of the GDPR in each paper: a) Complete, when the authors discuss a solution that involve the totality of the GDPR, b) Partial, when the authors deal with a specific issue addressed by the GDPR, i.e., checking privacy policy compliance, and c) Not Applicable (NA), when the GDPR is not involved in the paper. The fourth column *"Compliance"* indicates whether the authors of each paper present a mechanism to check compliance. The sixth column *"Available"* shows whether the solution proposed is publicly available. Finally, the seventh column *"Expert"* reports if any expert from the legal domain was involved in the realization of the research. We discuss the selected studies next.

Modeling the GDPR. There is some early work on conceptual modeling of the GDPR. In particular, Ayala-Rivera and Pasquale [3] propose a model-based approach to help organizations understand the data protection obligations imposed by the GDPR. Burmeister et al. [6] present an approach based on enterprise architecture (EA) models to help with checking GDPR compliance. The authors introduce a privacy-driven EA metamodel that provides recommendations to address GDPR concerns. Diamantopoulou et al. [11] present a GDPR-relevant metamodel for Privacy Level Agreements to support privacy management, based on analysis of privacy threats, vulnerabilities and trust relation-

Table 4 Related Work.

Reference	Year	GDPR Coverage	Compliance	Available	Expert
Ayala-Rivera and Pasquale [3]	2018	Partial	✓	✓	×
Burmeister et al. [6]	2019	Complete	×	✓	×
Diamantopoulou et al. [11]	2017	Partial	×	×	×
Sing [32]	2018	Complete	✓	✓	×
Caramujo et al. [8]	2019	Partial	×	✓	×
Pullonen and Matulevicius [27]	2019	Complete	×	✓	×
Tom et al. [38]	2018	Complete	×	✓	×
Chung et al. [9]	2008	NA	✓	✓	×
Panesar-Walawege et al. [26]	2013	NA	✓	✓	✓
Ranise and Siswantoro [28]	2020	Partial	✓	×	×
Guarda et al. [18]	2017	NA	✓	✓	✓
This article	**2021**	**Complete**	✓	✓	✓

ships in their Information Systems, whilst complying with laws and regulations. Sing [32] proposes a method based on a metamodel for analysing business processes of information systems and aligning them with the GDPR. Caramujo et al. [8] target privacy policies from the web and mobile applications and propose a domain-specific language along with model transformations for specifying privacy-policy models. Pullonen and Matulevicius [27] present a multi-level model to be used as an extension of the Business Process Model and Notation (BPMN) to enable the visualization, analysis, and communication of the privacy-policy characteristics of business processes. Tom et al. [38] present a preliminary GDPR model aimed at providing a simple, visual overview so that process implementers can better understand the associations between different entities in the GDPR. The authors describe an approach for using their proposed model as a tool to develop an organizational privacy policy along with an illustration of compliance-rule extraction. These existing strands of work either address narrow analytical use cases (e.g., only the compliance analysis of privacy policies) or focus on providing guidelines for the (manual) application of the GDPR. We go beyond the existing work by modeling the GDPR in a more holistic way and providing a systematic tailoring mechanism to support GDPR compliance automation in different contexts.

Checking Compliance. To the best of our knowledge, no automated approach for checking GDPR compliance has been published so far. However, there are a few threads of work that describe methodologies for assessing system compliance. Chung et al. [9] identify non-compliance issues in user-defined process models by matching these models against a standard model during both process specification and process execution. Panesar-Walawege et al. [26] propose a model-based approach to aid the suppliers of safety-critical systems in defining the evidence information necessary for certification according to standards and automatically detecting non-compliance issues in the collected evidence. Ranise and Siswantoro [29] devise an SMT-based tool for checking compliance of security policies at design time. They introduce an implementation that uses tools for policy analysis based on efficient Satisfiability Modulo Theories (SMT) solvers. Guarda et al. [18] propose a logic-based framework to support the specification of information system designs, purpose-aware access control policies, and legal requirements.

While being a useful source of inspiration, none of the above approaches can be directly adapted to the GDPR due to their main focus being different than data protection and privacy.

9 Conclusion

In this paper, we used UML and OCL to build a model-based representation of GDPR. The key motivation behind this research is to pave the way for the creation of automated, model-based GDPR compliance analysis solutions. Our research resulted in the development of a generic GDPR model alongside a detailed and well-defined strategy for specializing this model according to different contexts and for satisfying the requirements of different types of GDPR-related analysis. We presented several artifacts: (1) a GDPR conceptual model with full traceability to the GDPR, (2) a glossary to help explain the conceptual model, (3) 35 compliance rules extracted from GDPR and defined both in plain English and in OCL, and (4) a set of 20 GDPR variation points to specialize the generic model. Building on

the knowledge obtained from our modeling endeavor, we discussed several learned lessons. We also suggested potential strategies for solving the challenges we found in our work and promoting longer-term research focus on the model-based analysis of GDPR compliance.

In the future, we plan to work on the directions presented in Section 6 in order to enable a full realization of the approach outlined in Section 3. Furthermore, we will be working closely with legal experts on implementing a number of compliance analysis use cases, e.g., checking the compliance of data processing agreements with GDPR. Doing so will allow us to identify and address high-priority automation needs and help bridge the gap between software engineers and legal experts by developing more effective communication methods.

Acknowledgment

This paper was supported by Linklaters, Luxembourg's National Research Fund (FNR) under grant BRIDGES/19/IS/13759068/ARTAGO, and NSERC of Canada under the Discovery, Discovery Accelerator and CRC programs.

References

1. Alpaydin, E.: Machine Learning: The New AI. MIT Press (2016)
2. Arora, C., Sabetzadeh, M., Briand, L.C., Zimmer, F.: Extracting domain models from natural-language requirements: Approach and industrial evaluation. In: Proceedings of the 19th IEEE/ACM International Conference on Model Driven Engineering Languages and Systems (MoDELS'16), pp. 250–260 (2016)
3. Ayala-Rivera, V., Pasquale, L.: The grace period has ended: An approach to operationalize GDPR requirements. In: Proceedings of 31st IEEE International Conference on Requirements Engineering (RE'18), pp. 136–146 (2018)
4. Brambilla, M., Cabot, J., Wimmer, M.: Model-Driven Software Engineering in Practice, 2nd edn. Morgan & Claypool Publishers (2016)
5. Breaux, T.: Exercising due diligence in legal requirements acquisition: A tool-supported, frame-based approach. In: Proceedings of 17th IEEE International Conference on Requirements Engineering (RE'09), pp. 225–230 (2009)
6. Burmeister, F., Drews, P., Schirmer, I.: A privacy-driven enterprise architecture meta-model for supporting compliance with the general data protection regulation. In: T. Bui (ed.) 52nd Hawaii International Conference on System Sciences, HICSS 2019, Grand Wailea, Maui, Hawaii, USA, January 8-11, 2019, pp. 1–10. ScholarSpace (2019)
7. Cabot, J., Clarisó, R., Riera, D.: UMLtoCSP: A tool for the formal verification of UML/OCL models using constraint programming. In: Proceedings of the 22nd IEEE/ACM International Conference on Automated Software Engineering (ASE'07), pp. 547–548 (2007)
8. Caramujo, J., Rodrigues da Silva, A., Monfared, S., Ribeiro, A., Calado, P., Breaux, T.: RSL-IL4Privacy: A domain-specific language for the rigorous specification of privacy policies. Requirements Engineering **24**(1), 1–26 (2019)
9. Chung, P.W., Cheung, L.Y., Machin, C.H.: Compliance flow – Managing the compliance of dynamic and complex processes. Knowledge-Based Systems **21**(4), 332–354 (2008)
10. Clements, P., Northrop, L.: Software Product Lines: Practices and Patterns. Addison-Wesley (2001)
11. Diamantopoulou, V., Angelopoulos, K., Pavlidis, M., Mouratidis, H.: A metamodel for gdpr-based privacy level agreements. In: C. Cabanillas, S. España, S. Farshidi (eds.) Proceedings of the ER Forum 2017 and the ER 2017 Demo Track co-located with the 36th International Conference on Conceptual Modelling (ER 2017), Valencia, Spain, - November 6-9, 2017, *CEUR Workshop Proceedings*, vol. 1979, pp. 285–291. CEUR-WS.org (2017)
12. Emmerich, W., Finkelstein, A., Montangero, C., Antonelli, S., Armitage, S., Stevens, R.: Managing standards compliance. IEEE Transactions on Software Engineering **25**(6), 836–851 (1999)
13. EU-GDPR: EU GDPR portal (2019). URL https://eugdpr.org
14. European Union: The GDPR: New opportunities, new obligations. Justice and Consumers (2018)
15. European Union: General data protection regulation. Official Journal of the European Union (2018). URL http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679
16. France, R., Rumpe, B.: Model-driven development of complex software: A research roadmap. In: Proceedings of 2007 Workshop on the Future of Software Engineering (FOSE '07), pp. 37–54 (2007)
17. Ghanavati, S., Rifaut, A., Dubois, E., Amyot, D.: Goal-oriented compliance with multiple regulations. In: Proceedings of 22nd IEEE International Conference on Requirements Engineering (RE'14), pp. 73–82 (2014)
18. Guarda, P., Ranise, S., Siswantoro, H.: Security analysis and legal compliance checking for the design of privacy-friendly information systems. In: Proceedings o 22nd ACM on Symposium on Access Control Models and Technologies (SACMAT'17), pp. 247–254 (2017)
19. Hajri, I., Goknil, A., Briand, L.C., Stephany, T.: PUM-Conf: A tool to configure product specific use case and domain models in a product line. In: Proceedings of the 24th ACM SIGSOFT International Symposium on Foundations of Software Engineering (FSE'16), pp. 1008–1012 (2016)
20. Hajri, I., Göknil, A., Briand, L.C., Stephany, T.: Configuring use case models in product families. Software & Systems Modeling **17**(3), 939–971 (2018)
21. Ingolfo, S., Siena, A., Mylopoulos, J.: Nòmos 3: Reasoning about regulatory compliance of requirements. In: Proceedings of 22nd IEEE International Requirements Engineering Conference (RE'14), pp. 313–314 (2014)
22. van Lamsweerde, A.: Requirements Engineering - From System Goals to UML Models to Software Specifications. Wiley (2009)
23. Manning, C.D., Schütze, H.: Foundations of statistical natural language processing. MIT Press (2001)
24. OMG: Object Constraint Language - Version 2.4 (2017). URL https://www.omg.org/spec/OCL/2.4/PDF
25. OMG: Unified Modeling Language - Superstructure Version 2.5.1 (2017). URL https://www.omg.org/spec/UML/2.5.1/PDF

26. Panesar-Walawege, R.K., Sabetzadeh, M., Briand, L.C.: Supporting the verification of compliance to safety standards via model-driven engineering: Approach, tool-support and empirical validation. Information and Software Technology **55**(5), 836–864 (2013)
27. Pullonen, P., Tom, J., Matulevicius, R., Toots, A.: Privacy-enhanced BPMN: Enabling data privacy analysis in business processes models. Software & Systems Modeling pp. 1–30 (2019)
28. Rabinia, A., Ghanavati, S., Humphreys, L., Hahmann, T.: A methodology for implementing the formal legal-grl framework: A research preview. In: N. Madhavji, L. Pasquale, A. Ferrari, S. Gnesi (eds.) Requirements Engineering: Foundation for Software Quality, pp. 124–131. Springer International Publishing, Cham (2020)
29. Ranise, S., Siswantoro, H.: Automated legal compliance checking by security policy analysis. In: Computer Safety, Reliability, and Security (SAFECOMP'17 Workshops), pp. 361–372 (2017)
30. Sannier, N., Adedjouma, M., Sabetzadeh, M., Briand, L.C.: An automated framework for detection and resolution of cross references in legal texts. Requirements Engineering **22**(2), 215–237 (2017)
31. Shum, S.B., Hammond, N.: Argumentation-based design rationale: What use at what cost? International Journal of Human-Computer Studies **40**(4), 603–652 (1994)
32. Sing, E.: A meta-model driven method for establishing business process compliance to gdpr. Master's thesis, University of Tartu (2019)
33. Sleimi, A., Sannier, N., Sabetzadeh, M., Briand, L.C., Dann, J.: Automated extraction of semantic legal metadata using natural language processing. In: Proceedings of 26th IEEE International Requirements Engineering Conference (RE'18), pp. 124–135 (2018)
34. Soltana, G., Fourneret, E., Adedjouma, M., Sabetzadeh, M., Briand, L.C.: Using UML for modeling procedural legal rules: Approach and a study of luxembourg's tax law. In: J. Dingel, W. Schulte, I. Ramos, S. Abrahão, E. Insfrán (eds.) Model-Driven Engineering Languages and Systems - 17th International Conference, MODELS 2014, Valencia, Spain, September 28 - October 3, 2014. Proceedings, *Lecture Notes in Computer Science*, vol. 8767, pp. 450–466. Springer (2014)
35. Soltana, G., Sabetzadeh, M., Briand, L.C.: Practical model-driven data generation for system testing. arXiv preprint (arXiv:1902.00397) (2019). URL https://arxiv.org/pdf/1902.00397.pdf
36. Soltana, G., Sannier, N., Sabetzadeh, M., Briand, L.C.: Model-based simulation of legal policies: Framework, tool support, and validation. Software & Systems Modeling **17**(3), 851–883 (2018)
37. Tankard, C.: What the GDPR means for businesses. Network Security **6**, 5–8 (2016)
38. Tom, J., Sing, E., Matulevičius, R.: Conceptual representation of the GDPR: Model and application directions. In: Perspectives in Business Informatics Research, pp. 18–28 (2018)
39. Torre, D., Abualhaija, S., Sabetzadeh, M., Briand, L.C., Baetens, K., Goes, P., Forastie, S.: An AI-assisted approach for checking the completeness of privacy policies against GDPR. In: Proceedings of 28th IEEE International Conference on Requirements Engineering (RE'20) (2020)
40. Torre, D., Alferez, M., Soltana, G., Sabetzadeh, M., Briand, L.: Model Driven Engineering for Data Protection and Privacy: Application and Experience with GDPR – Appendix (2021). DOI 10.5281/zenodo.4564856. URL https://doi.org/10.5281/zenodo.4564856
41. Torre, D., Labiche, Y., Genero, M., Elaasar, M.: A systematic identification of consistency rules for UML diagrams. Journal of Systems and Software **144**, 121–142 (2018)
42. Torre, D., Soltana, G., Sabetzadeh, M., Briand, L.C., Auffinger, Y., Goes, P.: Using models to enable compliance checking against the GDPR: an experience report. In: 22nd ACM/IEEE International Conference on Model Driven Engineering Languages and Systems, MODELS 2019, Munich, Germany, September 15-20, 2019, pp. 1–11 (2019)
43. Zeni, N., Kiyavitskaya, N., Mich, L., Cordy, J.R., Mylopoulos, J.: GaiusT: Supporting the extraction of rights and obligations for regulatory compliance. Requirements Engineering **20**(1), 1–22 (2015)