A Comparison of System Description Models for Data Protection by Design

Pierre Dewitte
imec-CiTiP, KU Leuven
first.last@kuleuven.be

Kim Wuyts, Laurens Sion, Dimitri Van Landuyt
imec-DistriNet, KU Leuven
first.last@cs.kuleuven.be

Ivo Emanuilov, Peggy Valcke
imec-CiTiP, KU Leuven
first.last@kuleuven.be

Wouter Joosen
imec-DistriNet, KU Leuven
first.last@cs.kuleuven.be

ABSTRACT

Since the General Data Protection Regulation (GDPR) entered into force, every actor involved in the processing of personal data must comply with Data Protection by Design (DPbD). Doing so requires assessing the risks to data subjects' rights and freedoms and implementing appropriate countermeasures. While legal experts traditionally apply Data Protection Impact Assessments (DPIA), software engineers rely on threat modeling for their assessment.

Despite significant differences, both approaches nonetheless revolve around (i) a description of the system and (ii) the identification, assessment and mitigation of specific risks. In practice, however, DPIAs and threat modeling are usually performed in complete isolation, following their own, unharmonized lexicon and abstractions. Such as disconnect lowers the quality of the assessment and of the conceptual and architectural trade-offs

In this paper, we present (i) an overview of the legal and architectural modeling requirements and (ii) incentives and recommendations for aligning both modeling paradigms in order to support data protection by design from both a legal and a technical perspective.

CCS CONCEPTS

• Social and professional topics → Governmental regulations; **• Software and its engineering → Architecture description languages**; **System modeling languages**; **• Security and privacy →** *Security requirements*; *Software security engineering*;

KEYWORDS

data protection by design, privacy, threat modeling, system model

ACM Reference Format:
Pierre Dewitte, Kim Wuyts, Laurens Sion, Dimitri Van Landuyt, Ivo Emanuilov, Peggy Valcke, and Wouter Joosen. 2019. A Comparison of System Description Models, for Data Protection by Design. In *The 34th ACM/SIGAPP Symposium on Applied Computing (SAC '19), April 8–12, 2019, Limassol, Cyprus.* ACM, New York, NY, USA, Article 4, 4 pages. https://doi.org/10.1145/3297280.3297595

1 INTRODUCTION

Since the General Data Protection Regulation (GDPR) [17] entered into force, controllers are under the obligation to implement appropriate technical and organizational measures to ensure and demonstrate that personal data processing operations are performed in accordance with the Regulation [17, Art. 24(1)]. The GDPR adds that those measures should be implemented 'both at the time of the determination of the means for processing and at the time of the processing itself' [17, Article 25(1)] (Data Protection by Design (DPbD)). In other words, the GDPR allows controllers to tailor the extent of their compliance duty to the actual risks posed by their processing activities. This '*risk-based*' approach calls for an assessment of these risks and the adoption of mitigation strategies matching their potential level of harm for data subjects' rights and freedoms [19]. In addition, controllers must pro-actively embed those countermeasures into the architecture of their software systems and throughout the entire data processing life cycle.

While legal experts traditionally rely on Data Protection Impact Assessments (DPIA) [7] to do so, software engineers apply security and privacy threat modeling for their assessments. As a result, the representations of the system that serve as a starting point for both risk assessments differ significantly depending on the legal or technical focus of the exercise. Legal experts, on the one hand, describe the system using data protection-specific abstractions in order to streamline the evaluation of the proportionality and necessity of the processing activities. Software engineers, on the other, model the system as one or several views [23] of the technical architecture to analyze and address, amongst others, security and privacy risks by applying *threat modeling* [15, 20, 27, 28, 32]. Such a disconnect is a major stumbling block to interdisciplinary collaboration and impacts the overall quality of the compliance exercise.

In this paper, we explore and evaluate existing modeling paradigms for DPbD. Section 2 first outlines the requirements. Section 3 then provides a critical overview of the state-of-the-art in light of these requirements. Finally, Section 4 concludes the paper.

2 REQUIREMENTS

We define a set of requirements for system descriptions from two complementary perspectives: legal and architectural.

These requirements allow the assessment of (i) the expressivity of the system descriptions in terms of key concepts, and (ii) the degree to which the descriptions support the enforcements of relevant constraints; for example, for ensuring soundness, completeness, model quality.

SAC '19, April 8–12, 2019, Limassol, Cyprus

ACM ISBN 978-1-4503-5933-7/19/04.
https://doi.org/10.1145/3297280.3297595

2.1 Legal Description Requirements

Legal experts usually rely on DPIA. While the GDPR [17, Art. 35] only obliges controllers to conduct a DPIA for processing activities that are likely to result in a high risk to data subjects' rights and freedoms, such an approach also lays the groundwork for DPbD, an obligation which is (i) mandatory for every controller and (ii) based on a similar reasoning.

When it comes to the system representation, the GDPR [17, Art. 35(7)*a*] calls for a '*systematic description of the envisaged processing operations and the purposes of the processing*', but provides little details about the way the system should actually be represented in practice. The Article 29 Working Party (WP29) has published guidelines related to the execution and documentation of a DPIA [7]. More specifically, they list a series of criteria which controllers can use to assess the quality of a DPIA [7, Annex 2]. With regard to the above-mentioned 'systematic description', the WP29 specifies the elements that must be documented. In this paper, we only consider the requirements whose nature allows the representation within a system model. We excerpted the following legal concepts from the aforementioned guidelines:

Personal Data any information relating to an identified or identifiable natural person. [17, Art. 4(1)][7, Annex 2(1*a*, 1*b*, 2*a*)],

Data Subject the natural person whose data is being processed [17, Art. 4(1)][7, Annex 2(1*a*)]

Processing any operation performed on personal data [17, Art. 4(2)] [7, Annex 2(1*a*, 1*d*)]

Purpose the intent of the processing [17, Art. 5(1)*b*][7, Annex 2(1*a*, 2*a*)]

Lawful Ground the legal basis on which the processing is performed [17, Art. 5(1)*a* and 6][7, Annex 2(2*a*)]

Controller the natural or legal person which determines the purposes and means of the processing [17, Art. 4(7)][7, Annex 2(1*a*)]

Processor the natural or legal person which processes the personal data on behalf of the controller [17, Art. 4(8)][7, Annex 2(1*a*)]

Third Party the natural or legal person other than the data subject, controller, processor and persons who, under the direct authority of the controller or processor, are authorised to process personal data [17, Art. 4(10)][7, Annex 2(1*a*)]

Recipient the natural or legal person to which the personal data are disclosed [17, Art. 4(9)][7, Annex 2(1*b*)]

Representative the natural or legal person established in the Union who represents the controller or processor not established in the EU [17, Art. 4(17)][7, Annex 2(1*a*)]

Storage Period predefined period to store personal data after which it needs to be removed [17, Art. 5(1)*e*][7, Annex 2(1*b*, 2*a*)]

Assets Involved organizational and material infrastructure [7, Annex 2(1*d*)]

2.2 Architectural Description Requirements

Threat modeling is a well-known technique to elicit (security or privacy) threats in software systems. Examples of such methodologies are STRIDE for eliciting security threats [20, 27] and LINDDUN [15, 32] for eliciting privacy threats. Both methods start from a Data Flow Diagram-based (DFD) abstraction of the system to systematically elicit applicable security and privacy threats.

However, other representations could also be used. For this, Shostack [27] lists the following concepts that have to be included in a diagram for security threat modeling (based on Howard and Lipner [21]): **events** that drive the system, **processes** that are driven, **responses** each process generates and sends, **data sources** for each request and response, and **recipients** of each response.

Analysis of the data processed by the system is essential for threat elicitation, more so in the context of privacy than security. As such, we extend this list with an explicit notion of **data**. A final concept to be evaluated, which is relevant both from the legal and the software engineering perspective, is **tool support** for modeling.

2.3 Misalignment between the Requirements

This section briefly discusses the misalignment between the above two categories of requirements.

Data/Action Legal requirements focus exclusively on the processing of personal data while architectural requirements encompass all types of data, processing, and communications between the software elements.

Rationale Legal requirements include rationale-related concepts such as lawful grounds and purpose, something which is not supported by architectural requirements.

Actors The architectural requirements use broad concepts (e.g., sources or recipients) to model entities, including only entities that directly interact with the software system. Legal requirements, on the contrary, rely on specific categories of actors that are defined in the GDPR (e.g., controller, processor), which significantly impacts the allocation of responsibilities, and include actors that do not directly interact with the system (e.g., third parties).

Risk Because both approaches rely on their own concepts for risk assessment, misalignment of these concepts also leads to different approaches to risk assessment. The legal assessment is very broad, considering all risks to the data subjects' rights and freedoms, while the architectural assessment focuses exclusively on technical security and privacy risks.

Rather than creating a complex mapping between the legal and architectural requirements, we evaluate the support for them side-by-side. This makes the comparison easier and enables assessing the support from both perspectives separately.

3 STATE OF THE ART

In this section, we present overview of the modeling techniques and their support for describing the legal and architectural concepts listed above. For the evaluation, we apply the following scale:

0 not supported: The concept is not supported, or only partially because of the misalignment (Section 2.3).

1 limited, ad-hoc support: Only possible by (ab)using the support such as a free-format text with custom conventions.

2 supported: Possible in the model, but without explicit support for soundness checks, etc.

3 full support: Possible in the model including support for constraints, soundness checks, element relations, etc.

3.1 Architectural Approaches

The most common system description used in the context of security and privacy threat modeling are Data Flow Diagrams (DFDs) [14,

Table 1: Evaluation of existing modeling approaches w.r.t. legal and security/privacy architecture DPbD requirements.

Technique	Processing	Lawful grounds	Purpose	Personal Data	Data Subjects	Recipients	Controllers	Processors	Representatives	Third Parties	Storage Period	Assets	Events	Processes	Responses	Data Sources	Recipient	Data	Tooling
DFD [14, 20, 27]	●○○	○○○	○○○	●○○	○○○	●○○	○○○	○○○	○○○	●○○	●○○	○○○	●○○	●●●	●●○	●●●	●●●	●○○	●●●
DFD+dict.[14]	●○○	○○○	○○○	●●○	○○○	●○○	○○○	○○○	○○○	●○○	●○○	○○○	●○○	●●●	●●○	●●●	●●●	●●○	●○○
PA-DFD [6]	●○○	○○○	●●○	●●●	●●●	●○○	●●○	●●○	○○○	●○○	●●●	○○○	●○○	●●●	●●○	●●●	●●●	●●●	●○○
DFD+ontology [25]	●○○	○○○	●●●	●●●	●●●	●○○	●●●	●●●	○○○	●○○	●○○	●○○	●●○	●●●	●●○	●●●	●●●	●●●	●○○
CARiSMA ext. [2, 4]	●●●	○○○	●●●	●●○	○○○	●●●	○○○	○○○	○○○	●●○	●○○	○○○	●●○	●●○	●●○	●●○	●●○	●●○	●●●
petrinets [26]	●○○	●○○	●○○	●○○	●○○	●○○	●○○	●○○	●○○	●○○	●○○	●○○	●●○	●○○	●○○	●○○	●○○	●○○	●○○
ICN arch. [18]	○○○	○○○	○○○	●○○	○○○	●●○	●●○	○○○	○○○	○○○	○○○	○○○	●○○	●●○	●○○	●●○	●●○	○○○	●○○
DPIA methods [1, 8, 9, 11, 22, 29]	●●○	●●○	●●○	●●○	●●○	●●○	●●○	●●○	○○○	●●○	●●○	●●○	○○○	○○○	○○○	○○○	○○○	○○○	●●○
	Legal Concepts*												Architectural Concepts†						

Legend: ○ ○ ○: no support, ● ○ ○: limited, ad-hoc support, ● ● ○: supported, ● ● ●: full support including constraints, soundness checks, relations with other elements.

* *Legal concept requirements originate from the Article 29 Working Party requirements for a DPIA [7, Annex 2]*

† *Architectural concept requirements originate from Shostack's requirements for threat modeling diagrams [27]*

15, 20, 27, 32]. Both STRIDE [20, 27], for security threats, and LINDDUN [15, 32], for privacy threats, rely on this type of system description. Its simplicity (a DFD consists of only 5 element types) allows collaboration among stakeholders with different backgrounds (e.g., technical, business, legal, etc.). The Microsoft threat modeling tool [13] provides tool support for this activity and includes a number of soundness checks. As Table 1 shows, the DFD-based approaches provide little support for describing the necessary legal concepts. The type of support in these diagrams remains limited to the inclusion of only those recipients, third parties, or data subjects that actually directly interact with the system as a source or recipient. Furthermore, such a mapping remains implicit or requires some ad-hoc annotations on the elements. The same observation holds for processes (the technical realization) which do not cover the full scope of the legal concept processing.

The basic DFDs used in threat modeling approaches do not document the data in the system either. It is only documented implicitly via the flows. Data dictionaries [14] do separately document the involved data types. Although introduced together with the DFD, it is not used in the context of threat modeling.

Antignac et al. [6] have extended the DFD modeling notation to include data protection specific concepts. They include several legal concepts (such as purpose, personal data, and storage period), but do not support lawful grounds, representatives, and assets.

Oliver [25] extends DFDs with an ontological approach to explicitly support a number of legal concepts. It does lack support for lawful grounds and does not have (publicly available) tool support.

Ahmadian et al. [2, 4] propose a UML extension for privacy that includes stereotypes for sensitiveData, granularity, objectives and ABAC (attribute based access control) and privacy preferences purpose, visibility, granularity, and retention. Tool support is available with the integration in the CARiSMA [3] tool.

Petrinets can also be used to model and analyze privacy by design [16, 26]. They can be used to model the business flows and algorithms within processes, similar to an activity diagram. They can, at least in theory, also be used to model the system as a whole. Given their limited building blocks (i.e. *states* and *transitions* of a process, connected by *arcs*), there is no formal definition or possibility to make certain concepts mandatory in the system.

Fotiou et al. [18] propose the use of ICN architectures as input for a privacy analysis. An ICN network consists of *data owners* (which are in control of the data, and hence map to 'controller'), *consumers* (which correspond with 'recipients'), *storage nodes* (which correspond with 'data sources'), *resolvers* (which correspond with 'processes'), and two information containers: *data flows* and *data pools*.

3.2 Modeling Paradigm used in DPIAs

A DPIA always starts with an extensive description of all the personal data processing activities, as well as the identification and qualification of the actors involved. It is usually followed by: (i) the identification and documentation of data protection threats, (ii) the implementation of appropriate technical and organizational measures, (iii) the documentation of the process to ensure controller accountability, and (iv) a periodic monitoring and review phase.

Most of the time, the said description is performed by documenting the required legal concepts listed in Section 2.1. Guidance from national supervisory authorities is, however, limited to high-level advices, non-binding table templates, and knowledge bases [1, 8, 9, 11, 22, 29]. As a result, this exercise is usually performed manually, which requires tremendous efforts, can lead to human errors, and is highly sensitive to changes in the system. Only CNIL provides a tool to aid the documentation and assessment phase [12]. Given its legal focus, such an assessment also overlooks the technical aspects that are traditionally addressed in threat modeling but are nonetheless relevant to ensure compliance with data protection rules such as, for example, security obligations. Given the misalignment of requirements (as highlighted in Section 2.3), architectural concepts are only partially covered (evaluated as '0' in Table 1). For example, a legal recipient is someone external to the system who receives personal data, while the recipient concept from an architectural point corresponds with the receiving side of each interaction between two system components.

4 CONCLUSION

In this paper, we gathered an overview of the requirements imposed by the GDPR [17] and as laid out by the WP29 [7, Annex 2]. We have complemented these legal requirements with security and privacy architecture requirements derived from threat modeling best practices. Using this set of requirements, we assessed to which degree existing approaches support the necessary modeling abstractions for performing a comprehensive risk assessment.

We observe a clear dichotomy between architectural and legal approaches for data protection by design. They show inherently different aims targeting either legal or architectural description support. None of the approaches, however, provide support for a comprehensive description in both legal and architectural concepts. There are, however, strong incentives to integrate both views.

First, implementing the measures mandated by the GDPR [17, Art. 24(1) and 25(1)] often requires a technical insight into the system. This is notably the case for the obligation to guarantee the security of processing operations, an assessment which lies at the heart of threat modeling but is insufficiently supported in traditional DPIA methodologies [17, Art. 5(1)*f* and 32]. Similarly, an in-depth representation of all the data flows in a software system facilitates the identification of processing operations that might be overlooked by legal experts and highlights, for example, the need to specify a lawful ground [17, Art. 5(1)*a* and 6] or perform a compatibility assessment [17, Art. 5(1)*b*].

Second, it would drastically simplify the compliance exercise by matching architectural abstractions with their legal counterpart. Modeling, for example, a controller or a processor rather than an external entity already hints at the allocation of (i) responsibility for compliance, (ii) accountability for the measures implemented, and (iii) liability in case of non-compliance. Similarly, enriching representations with GDPR-specific concepts such as purpose and lawful ground would allow for the early identification and mitigation of data protection concerns such as compliance with the lawfulness [17, Art. 5(1)*a* and 6], purpose limitation [17, Art. 5(1)*b*], and data minimization [17, Art. 5(1)*c*].

Finally, aligning both modeling paradigms would ensure the consistency and validity of the countermeasures over time, since changes brought to the system representation will impact and orient the continuous compliance effort. Modeling an additional processor will, for instance, raise the need to draft a contract in accordance [17, Art. 28]. Not only will this strengthen the relevance of the technical and organizational trade-offs made during the design stage, but it will also pave the way for the automation of some of the requirements stemming from the GDPR on the basis of an accurate, up-to date representation of the system.

Acknowledgments

The research was partially funded by the Research Fund KU Leuven and the PRiSE research project.

REFERENCES

[1] Agencia Española de protección de datos (AEPD). 2018. Guía práctica para las Evaluaciones de Impacto en la Protección de los datos sujetas al RGPD. https://iapp.org/media/pdf/resource_center/Guia_EvaluacionesImpacto.pdf

[2] Amir Shayan Ahmadian, Jan Jürjens, and Daniel Strüber. 2018. Extending Model-Based Privacy Analysis for the Industrial Data Space by Exploiting Privacy Level Agreements. In *Proceedings of ACM SAC 2018: Privacy by Design in Practice.*

[3] Amir Shayan Ahmadian, Sven Peldszus, Qusai Ramadan, and Jan Jürjens. 2017. Model-based Privacy and Security Analysis with CARiSMA. In *Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering (ESEC/FSE 2017).*

[4] Amir Shayan Ahmadian, Daniel Strüber, Volker Riediger, and Jan Jürjens. 2018. Supporting privacy impact assessment by model-based privacy analysis. In *Proceedings of ACM SAC 2018: Software Engineering.*

[5] Rehab Alnemr, Erdal Cayirci, Lorenzo Dalla Corte, Alexandr Garaga, Ronald Leenes, Rodney Mhungu, Siani Pearson, Chris Reed, Anderson Santana de Oliveira, Dimitra Stefanatou, Katerina Tetrimida, and Asma Vranaki. 2016. A Data Protection Impact Assessment Methodology for Cloud. In *Privacy Technologies and Policy.* Springer.

[6] Thibaud Antignac, Riccardo Scandariato, and Gerardo Schneider. 2016. A privacy-aware conceptual model for handling personal data. 9952 LNCS (2016).

[7] Article 29 Data Protection Working Party. 2017. Guidelines on Data Protection Impact Assessment (DPIA) and determining whether processing is "likely to result in a high risk" for the purposes of Regulation 2016/679 (WP248 rev.01). (2017).

[8] Authorité de Protection des Données (APD). 2018. Modèle de registre des activités de traitement. https://www.autoriteprotectiondonnees.be/canevas-de-registre-des-activites-de-traitement

[9] Authorité de Protection des Données (APD). 2018. Recommandation n° 01/2018 du 28 février 2018 concernant l'analyse d'impact relative à la protection des données et la consultation préalable. https://www.autoriteprotectiondonnees.be/sites/privacycommission/files/documents/recommandation_01_2018.pdf

[10] Felix Bieker, Michael Friedewald, Marit Hansen, Hannah Obersteller, and Martin Rost. 2016. A Process for Data Protection Impact Assessment Under the European General Data Protection Regulation. In *Privacy Technologies and Policy.* Springer.

[11] Commission Nationale de l'Informatique et des Libertés. 2018. Privacy Impact Assessment (PIA). https://www.cnil.fr/en/privacy-impact-assessment-pia

[12] Commission Nationale de l'Informatique et des Libertés. 2018. The open source PIA software helps to carry out data protection impact assesment. https://www.cnil.fr/en/open-source-pia-software-helps-carry-out-data-protection-impact-assesment

[13] Microsoft Corporation. 2016. Microsoft Threat Modeling Tool 2016. https://www.microsoft.com/en-us/download/details.aspx?id=49168

[14] Tom DeMarco. 1979. *Structured Analysis and System Specification.* Yourdon Press.

[15] Mina Deng, Kim Wuyts, Riccardo Scandariato, Bart Preneel, and Wouter Joosen. 2011. A privacy threat analysis framework: supporting the elicitation and fulfillment of privacy requirements. *Requirements Engineering* 16, 1 (2011), 3–32.

[16] Laurence Diver and Burkhard Schafer. 2017. Opening the black box: Petri nets and Privacy by Design. *International Review of Law, Computers & Technology* (2017).

[17] European Union. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016. OJ L 119, 04.05.2016, p. 1–88. *Official Journal of the European Union* 59, L 119 (may 2016), 1–88.

[18] Nikos Fotiou, Somaya Arianfar, Mikko Särelä, and George C. Polyzos. 2014. A Framework for Privacy Analysis of ICN Architectures. In *Privacy Technologies and Policy.* Springer International Publishing, Cham, 117–132.

[19] Raphael Gellert. 2018. Understanding the notion of risk in the General Data Protection Regulation. *Computer Law & Security Review* 34, 2 (2018), 279 – 288.

[20] Michael Howard and Steve Lipner. 2006. *The Security Development Lifecycle.*

[21] Michael Howard and Steve Lipner. 2009. *Writing Secure Code, Second Edition.*

[22] International Commissioner's Office (ICO). 2018. How do we carry out a DPIA? https://ico.org.uk/for-organisations/guide-to-the-general-data-protection-regulation-gdpr/data-protection-impact-assessments-dpias/how-do-we-carry-out-a-dpia/

[23] Philippe B Kruchten. 1995. The 4+ 1 view model of architecture. *IEEE software* 12, 6 (1995), 42–50.

[24] Marie Caroline Oetzel and Sarah Spiekermann. 2014. A systematic methodology for privacy impact assessments: A design science approach. *European Journal of Information Systems* 23, 2 (2014), 126–150.

[25] Ian Oliver. 2014. *Privacy engineering: A dataflow and ontological approach.*

[26] Mushfiqur Rahman. 2017. A Petri Nets Semantics for Privacy-Aware Data Flow Diagrams. (2017).

[27] Adam Shostack. 2014. *Threat Modeling: Designing for Security.*

[28] Laurens Sion, Koen Yskout, Dimitri Van Landuyt, and Wouter Joosen. 2018. Risk-based design security analysis. *2018 IEEE/ACM First International Workshop on Security Awareness from Design to Deployment (SEAD18).*

[29] Unabhängiges Landeszentrum für Datenschutz (ULD). 2017. The Standard Data Protection Model: A concept for inspection and consultation on the basis of unified protection goals. https://www.datenschutzzentrum.de/uploads/sdm/SDM-Methodology_V1.0.pdf

[30] Kush Wadhwa and Rowena Rodrigues. 2013. Evaluating privacy impact assessments. *Innovation: The European Journal of Social Science Research* (2013).

[31] David Wright. 2013. Making Privacy Impact Assessment More Effective. *The Information Society* 29, 5 (2013), 307–315.

[32] Kim Wuyts. 2015. *Privacy Threats in Software Architectures.* Ph.D. Dissertation. KU Leuven.