NLP-based Automated Compliance Checking of Data Processing Agreements against GDPR

Orlando Amaral, Member, IEEE, Muhammad Ilyas Azeem, Member, IEEE, Sallam Abualhaija, Member, IEEE, and Lionel C Briand, Fellow, IEEE

Abstract—When the entity processing personal data (the processor) differs from the one collecting personal data (the controller), processing personal data is regulated in Europe by the General Data Protection Regulation (GDPR) through data processing agreements (DPAs). Checking the compliance of DPAs contributes to the compliance verification of software systems as DPAs are an important source of requirements for software development involving the processing of personal data. However, manually checking whether a given DPA complies with GDPR is challenging as it requires significant time and effort for understanding and identifying DPA-relevant compliance requirements in GDPR and then verifying these requirements in the DPA. Legal texts introduce additional complexity due to convoluted language and inherent ambiguity leading to potential misunderstandings. In this paper, we propose an automated solution to check the compliance of a given DPA against GDPR. In close interaction with legal experts, we first built two artifacts: (i) the “shall” requirements extracted from the GDPR provisions relevant to DPA compliance and (ii) a glossary table defining the legal concepts in the requirements. Then, we developed an automated solution that leverages natural language processing (NLP) technologies to check the compliance of a given DPA against these “shall” requirements. Specifically, our approach automatically generates phrasal-level representations for the textual content of the DPA and compares them against predefined representations of the “shall” requirements. By comparing these two representations, the approach not only assesses whether the DPA is GDPR compliant but it further provides recommendations about missing information in the DPA. Over a dataset of 30 actual DPAs, the approach correctly finds 618 out of 750 genuine violations while raising 76 false violations, and further correctly identifies 524 satisfied requirements. The approach has thus an average precision of 89.1%, a recall of 82.4%, and an accuracy of 84.6%. Compared to a baseline that relies on off-the-shelf NLP tools, our approach provides an average accuracy gain of ≈20 percentage points. The accuracy of our approach can be improved to ≈94% with limited manual verification effort.

Index Terms—Requirements Engineering (RE), The General Data Protection Regulation (GDPR), Regulatory Compliance, Natural Language Processing (NLP), Data Processing Agreement (DPA), Privacy.

1 INTRODUCTION

The advances in AI technologies led to increasingly integrated modern software systems (e.g., mobile applications) in everyday life activities. Applications are often developed to provide convenient services or assistance to individuals (e.g., Amazon Alexa). However, they also entail challenges regarding compliance to privacy standards and data protection requirements, e.g., when sharing personal data with third parties [1]. Individuals (i.e., the applications’ end-users) who provide explicit consent to these applications for allowing the processing of their personal data must have guarantees that their data remains protected, also when shared with third parties [2]. In Europe and even worldwide, the general data protection regulation (GDPR) was put into effect in 2018 to oblige organizations to provide such data protection guarantees when handling personal data of European residents. Violating GDPR can levy penalty fines reaching up to tens of millions of euros [3]. Software systems which typically involve processing or sharing personal data are subject to compliance with GDPR. Developing GDPR-compliant software requires taking into account the legally-binding agreements signed between different organizations involved in collecting and processing personal data. Examples of such agreements include privacy policies between organizations and individuals (often signed by individuals) as well as data processing agreements (DPAs) between two organizations involved in collecting and processing personal data. Individuals are not necessarily aware of DPAs. We focus in our work on verifying the compliance of DPAs against the provisions of GDPR. From a requirements engineering (RE) standpoint, this compliance verification is a prerequisite for capturing a complete set of privacy-related requirements covered in the DPAs to be implemented in software systems for processing personal data.

Data processing normally involves an individual (i.e., data subject) who willingly shares personal data, an organization (i.e., data controller) that collects and in many cases further shares personal data, and another organization (i.e., data processor) that processes personal data on behalf of the controller. A data processor can share personal data again with yet another organization (i.e., sub-processor) to perform
some data processing services on its behalf. Individuals in this chain are often aware of the terms based on which their personal data is collected and handled as described in privacy notices of the data controller. All subsequent sharing of individual’s personal data with processors (and sub-processors) is however not directly visible to individuals. According to GDPR, both the controller and processor should share the responsibility of protecting personal data of individuals. Consequently, a DPA listing privacy-related requirements should be established between the controller and the data processor(s) [4].

DPAs are legally binding agreements that regulate the data processing activities according to GDPR. To be deemed GDPR-compliant, the DPA must explicitly cover all the criteria imposed by the GDPR provisions concerning data processing. Establishing a DPA includes setting terms for how data is used, stored, protected, and accessed as well as defining the obligations and rights of the controller and processor. By signing a DPA, the processor is obliged to ensure that any software system deployed for processing personal data has to also comply with GDPR. To illustrate, consider the scenario of an educational institution (called as Sefer University) which shares personal data of its employees with a third-party service provider (e.g., accounting office) for a particular purpose (e.g., payroll administration). Examples of shared personal data include: first and last name, birth date, marital status, annual summary of leave or sickness absences and a scanned copy of a valid personal identification. The data controller in this scenario is Sefer University which collects the personal data of the employees (data subjects). The data processor is the service provider (in our example, the accounting office) performing financial services on behalf of the controller. Details concerning this service provider (e.g., the name and address of the hired accounting office) might not be shared with employees. To ensure that their personal data is sufficiently protected, GDPR requires Sefer university to sign a DPA with the accounting office.

1.1 Running Example

Fig. 1 shows on the left-hand side an excerpt from a DPA that addresses the scenario above, and on the right-hand side a set of DPA-related requirements extracted from GDPR, ordered by their occurrence in the DPA. In Section 3 we elaborate on the complete list of GDPR requirements. In the remainder of the paper, we use requirements to refer to the privacy-related requirements extracted from GDPR concerning data processing, and statements to refer to the textual content of a given DPA. For simplicity, a statement is regarded as equivalent to one sentence in the DPA.

Checking the compliance of the DPA in Fig. 1 requires mapping the statements (marked as S1 – S13) in the DPA against the requirements envisaged by GDPR (marked as R1 – R7, R9, R12 – R16, and R19). Note that the requirements numbering is not sequential and is defined by the complete list presented in Table 1. The most immediate straight-forward solution that comes to mind for matching statements with requirements is by applying semantic similarity measures, e.g., cosine similarity. In brief, cosine similarity is a measure that assigns a score between 0 and 1 to a pair of text sequences. The higher the score, the more semantically similar the two text sequences. Following this, a statement is satisfying a given requirement when the cosine similarity is sufficiently high. Below, we discuss the limitations of such a simple solution.

The example shows that S1 satisfies both R1 and R2 since it clearly identifies the data controller (Sefer University) and processor (Levico Accounting GmbH). However, computing the semantic similarity between S1 and R1 and R2 results in scores equal to 0.02 and 0.14, respectively. Low similarity scores in this case should be expected since S1 contains specific entities in place of the identity and contact details of the controller and processor (required by GDPR). Without considering S2 and S3, one cannot distinguish between the data controller and processor fully automatically. Since this is key information that has a significant impact on compliance checking, we assume in our work that the name(s) of the data controller and processor are provided as input by the human analyst. To properly verify the compliance of a given DPA, the requirements have to be checked for the right entity. For example, the processor’s obligations in R7, R9, R12-R16, and R19 in Fig. 1 must be associated with the processor. On a similar basis, S12 has a similarity score of 0.3 with R3, S6 0.38 with R4, S7 0.53 with R5 and 0.33 with R6. All these semantic similarity scores are too low to enable a decision about compliance. Using semantic similarity is not suitable for checking the compliance against R1 – R6 since they require content that varies across DPAs, e.g., a list of collected personal data or the processing purpose.

The example further shows that S8 satisfies neither R7 nor R9, yet it has a semantic similarity of 0.52 and 0.57, respectively. S8 is about having a written agreement with sub-processors, while R7 concerns acquiring a written approval from the controller before engaging sub-processors. R9 is related to processing personal data only based on documented instructions from the controller. Though there is some resemblance in the wording of S8 and that of R7 and R9, it is still clear that S8 does not comply with either requirements.

The example illustrates that statements in DPAs are often verbose, i.e., include more content that is needed for GDPR compliance. For example, S9 satisfies R13 with a similarity score of 0.87, whereas S10 satisfies both R12 and R14 with a similarity score of 0.52. This indicates that irrelevant content in S10 reduces the overall semantic similarity, and thus impedes compliance verification. Computing the similarity with only the relevant part of S10 (shaded yellow) leads to higher scores: 0.71 with R12 and 0.60 with R14.

Further, statements can satisfy multiple requirements. For example, S11 satisfies R15 and R16. The similarity scores between S11 and R15 and R16 are 0.82 and 0.79, respectively. However, S11 does not specify stating to whom data breaches shall be notified (that is the supervisory authority in R15 or the data subject in R16). In this case, S11 can be regarded as partially satisfying both requirements. We see that a GDPR requirement can contain different key elements which must be verified in the DPA statement. For instance, key elements in R15 include the processor, assist in notifying, the controller, personal data breach, and supervisory authority.

Verifying all key elements is necessary to understand the missing content in a statement. To become compliant, S11
The above discussion highlights that the seemingly simple task of mapping statements in a given DPA to GDPR requirements cannot be addressed accurately through straightforward automation based on semantic similarity. Alternatively, manually analyzing the content of DPAs for compliance against GDPR is a very tedious and time-consuming task for legal experts and requires engineers alike. In this paper, we propose a natural language processing (NLP)-based automation for checking the compliance of DPAs against GDPR. We develop our solution in close collaboration with legal experts from our industry partner (Linklaters LLP - a multinational law firm). Linklaters has top-tier rankings across many legal-practice areas, including funds investment, banking, and finance. The legal experts in Linklaters have long experience in providing GDPR advisory and compliance services.

Ensuring that the DPA excerpt in Fig. 1 is compliant with GDPR, and consequently provides a complete list of requirements, has an impact on building a GDPR-compliant software system. For example, a system deployed by the processor shall not process or store any personal data outside the list specified in S7. The system shall further implement: (i) appropriate protection procedures (e.g., encryption, data anonymization) to ensure the required level of security in accordance with S10, and (ii) a data breach notification procedure, corresponding to S11, which shall send a text message to the controller and also collect an acknowledgment of receipt.

As we elaborate in Section 5 much work has been done in the RE literature on automating regulatory compliance. Despite being an essential source for compliance requirements of data processing activities in software systems, automated compliance checking of DPAs against GDPR has not been previously studied in RE. Moreover, none of the existing work leverages the key elements of requirements in
GDPR (e.g., the case of R15 and R16 discussed above) for a phrasal-level, automated compliance checking. Analyzing the content of DPAs at a phrasal level is beneficial for understanding the exact missing information content in the DPA and hence recommending remedies to prevent non-compliance with GDPR.

In our work, we propose using semantic frames for generating an intermediate representation to characterize the information content in each GDPR requirement and thus enable a meaningful, phrasal-level decomposition of that requirement. As we elaborate in Section 2, a semantic frame (SF) of a given sentence describes an event and a set of participants, where each participant has a specific semantic role (SR) in that event. Fig. 2 illustrates SF elements describing a processing event in a requirement example extracted from GDPR Art. 28(3)(a) – corresponding to R9 in Fig. 1. The event processing, identified through the SR action, evokes three participants: “the processor” who performs processing (i.e., actor), “personal data” on which processing is performed (i.e., object) and “only on documented instructions from the controller” that restricts processing (i.e., constraint).

1.2 Research Questions

This paper investigates four Research Questions (RQs):

RQ1: What are the requirements for checking the compliance of a DPA according to GDPR? We answer RQ1 by defining a set of 45 “shall” requirements concerning the compliance of a DPA. Verifying these requirements in DPAs ensures that the DPAs contain a complete set of necessary legal requirements, to be implemented in software systems, concerning data processing activities. We further create a glossary table that defines the legal concepts related to DPA compliance and provide traceable links to GDPR articles. We present the compliance requirements in Section 3.

RQ2: How can a DPA be automatically checked for compliance against GDPR at a phrasal level? To address RQ2, we first manually define SF-based representations for the 45 compliance requirements from RQ1. Then, we devise an automated approach that leverages various NLP technologies to automatically generate SF-based representations for the textual content of a given DPA. Our approach then verifies the compliance of the DPA by aligning its respective SF-based representations against the representation of each GDPR requirement. We elaborate our approach in Section 4.

RQ3: How accurately can we check the compliance of a given DPA? RQ3 assesses the accuracy of our approach in checking the compliance of DPAs according to GDPR provisions. Over an evaluation set of 30 real DPAs, our approach successfully detects 618 out of 750 violations and introduces false violations in 76 cases. Our approach has thus a precision of 89.1% and a recall of 82.4%. Compared to a baseline that employs off-the-shelf NLP tools, our approach yields on average a gain of ≈20 percentage points in accuracy, and ≈17 percentage points in both precision and recall. We discuss RQ3 and the subsequent RQs in Section 5.

RQ4: How efficient is our approach in terms of execution time? Based on our experience, reviewing entirely manually the content of a given DPA containing about 200 statements takes on average 30 minutes if one is intimately familiar with the GDPR articles relevant to data processing activities. In contrast, our approach automatically analyzes a DPA of the same size in ≈2.5 minutes, and further produces as output a detailed report which summarizes the results, including the compliance decision together with the missing information and the respective recommendations to achieve compliance.

1.3 Contributions

We take steps toward addressing the limitations outlined above. Concretely, our contributions are as follows:

1. We extract in close interaction with subject-matter experts, a list of 45 compliance requirements from the GDPR provisions concerning data processing. We further simplify and document these compliance requirements as “shall” requirements. Our motivation is to increase readability for requirements engineers by using a familiar format, while maintaining usability for legal experts since these requirements rely on GDPR terminology.

2. We devise an automated approach that leverages NLP technologies for automatically checking the compliance of DPAs against GDPR. More specifically, we rely on syntactic and semantic parsing to generate intermediate SF-based representations that are then compared for compliance checking. Our proposed SF-based representation is based on a set of semantic roles with intuitive descriptions. Future changes in compliance requirements is a common concern for legal experts and requirements engineers. With enough training, additional compliance requirements can be added to our automated approach by defining their respective SF-based representations. The SF-based representation is used as an enabler for automated compliance checking.

3. We empirically evaluate our approach on 54 real DPAs. We randomly split the DPAs into two mutually exclusive subsets as follows. The first subset containing 24 DPAs was used for creating the SF-based representations and developing our approach, whereas the second subset, including the remaining 30 DPAs, was used exclusively for evaluation. The evaluation set collectively contains a total of 7048 statements manually reviewed and analyzed for compliance by third-party annotators. On this evaluation set, our approach yields a precision of 89.1%, and a recall of 82.4%. Our approach outperforms a straight-forward baseline that employs NLP off-the-shelf tools with a percentage point gain of ≈17 in both precision and recall.

4. We propose computing and assigning scores to requirements found to be satisfied in a given DPA. Such score indicates the maximum matching degree between a requirement and any DPA statement satisfying it. The matching degree represents the ratio of key elements that are present in a DPA statement compared to the elements to be expected given the requirement. We discuss how matching degrees...
are computed in detail in Section 3. These scores are intended to facilitate the review and validation of findings from our automated approach by a human analyst. For example, our empirical evaluation shows that the analyst can improve the overall accuracy of our automated approach by 6.4 percentage points when reviewing statements satisfying requirements with the maximum matching degree scores as long as the scores are less than or equal to a threshold of 0.5. The proportion of these statements, on average, is \( \approx 6\% \) (corresponding to an average of 12 statements in a DPA). This accuracy improvement comes with a gain of 1.2 and 11.5 percentage points in precision and recall, respectively.

1.4 Structure.

Section 2 explains the necessary background for our approach. Section 3 presents how we extracted the DPA-related requirements from GDPR. Section 4 describes our automated approach. Section 5 evaluates our approach. Section 6 reviews related work. Section 7 discusses threats to validity. Section 8 describes how our methodology can be adapted beyond GDPR or DPAs. Finally, Section 9 concludes the paper.

2 BACKGROUND

In this section, we briefly introduce GDPR, and explain the NLP background on semantic frames.

2.1 GDPR

GDPR [6] is the benchmark for data privacy regulations in Europe with 173 recitals and 99 articles divided into 11 chapters. GDPR applies primarily to organizations in Europe, but it can also impose obligations on organizations outside Europe when they offer goods or services, or monitor individuals in Europe. According to GDPR, there are often two main actors that should be identified in agreements related to processing personal data, namely a data controller and data processor. A data controller determines the purpose and means of processing personal data, whereas a data processor processes personal data based on the instructions of the controller. GDPR imposes various obligations on a given organization when it is acting as a processor or controller, the latter being subject to more provisions than the former.

Data processors are obliged to: (1) implement adequate technical and organizational measures to ensure the security of personal data, and, in case of data breaches, notify the controllers without undue delay; (2) appoint a statutory data protection officer (if needed) and conduct a formal data protection impact assessment for certain types of high-risk processing; (3) keep records about their data processing; and (4) comply to GDPR restrictions when transferring personal data outside Europe. Data controllers have to meet the obligations listed above, and further: (1) adhere to six core personal data processing principles, namely, fair and lawful processing, purpose limitation, data minimization, data accuracy, storage limitation, and data security; (2) keep identifiable individuals informed about how their personal data will be used; and (3) ensure the individual rights envisaged by GDPR, e.g., the right to be forgotten and the right to lodge a complaint.

GDPR includes some specific provisions in relation to DPAs, which in turn can have an impact on software development. For example, GDPR requires data processors to return or delete personal data without undue delay after the processing service is completed. For some cloud service providers, implementing this requirement, where immediate deletion is not possible, requires archiving personal data first until the next delete command is performed. We elaborate the GDPR requirements concerning DPAs in Section 3.

2.2 Semantic Frames

A semantic frame (SF) describes a set of elements that participate in a particular event [5]. Each element in the frame is a text span that is labeled with a semantic role (SR), denoted as [span]role. For instance, the SF about a purchase event in the sentence “[William]buyer purchased [a brand new car]item” contains the SRs buyer and item and spans “William” and “a brand new car”, respectively. SFs are often represented as predicate-arguments structure, where a predicate describes an event represented by the text span of the verb, and each argument describes a participant in the event labeled with a specific SR. A predicate evokes a set of expected arguments based on the event. For example, the above sentence contains the arguments buyer and purchased item, but it could also contain seller and price. The NLP community has long been creating machine-readable lexical resources like FrameNet [7] and PropBank [8] that contain manually-labeled predicates and arguments.

One way to identify the SF of a given sentence is through semantic role labeling (SRL), which is the task of identifying the different SRs occurring in the sentence [9]. In SRL, one identifies the text spans that provide answers to questions about who (i.e., actor) did what (i.e., action) to whom (beneficiary), when (i.e., time) and why (i.e., reason) [10], [11]. SRL can be employed to improve the solutions of many downstream NLP tasks such as machine translation [12], question answering [13], [14] and relation extraction [15], [16]. For instance, question answering can be improved by aligning the SRs in the question with those in the likely answer. If we identify the SRs in the question “[What]item did [Mazda]seller [sell]action to [William]buyer?”, extracting the answer from the above example (i.e., “a brand new car”) becomes straightforward. Inspired by the NLP literature [14], [17], [18], we check the compliance of a given DPA against GDPR requirements by generating SF-based representations. In particular, we use the SR action to define the predicate of an SF, and a set of other SRs (e.g., actor and reason) to define the arguments as we elaborate in Section 3.

3 COMPLIANCE REQUIREMENTS FOR DATA PROCESSING IN GDPR (RQ1)

In this section, we present two artifacts to answer RQ1, namely (1) a total of 45 compliance requirements for DPA according to GDPR; and (2) a glossary table defining the legal concepts in these requirements, including traceability to GDPR articles. The artifacts are provided in Table 1 as well as in the tables provided as supplementary material.
### TABLE 1: Mandatory DPA Compliance Requirements in GDPR.

| ID (Cat\(^1\)) | Requirement (Reference\(^2\)) |
|----------------|--------------------------------|
| R1 (MD1)       | The DPA shall contain at least one controller’s identity and contact details. (Linklaters LLP) |
| R2 (MD2)       | The DPA shall contain at least one processor’s identity and contact details. (Linklaters LLP) |
| R3 (MD3)       | The DPA shall contain the duration of the processing. (Art. 28(3)) |
| R4 (MD4)       | The DPA shall contain the nature and purpose of the processing. (Art. 28(3)) |
| R5 (MD5)       | The DPA shall contain the types of personal data. (Art. 28(3)) |
| R6 (MD6)       | The DPA shall contain the categories of data subjects. (Art. 28(3)) |
| R7 (PO1)       | The processor shall not engage a sub-processor without a prior specific or general written authorization of the controller. (Art. 28(2)) |
| R8 (PO2)       | In case of general written authorization, the processor shall inform the controller of any intended changes concerning the addition or replacement of sub-processors. (Art. 28(2)) |
| R9 (PO3)       | The processor shall process personal data only on documented instructions from the controller. (Art. 28(3)(a)) |
| R10 (PO4)      | If the processor requires by Union or Member State law to process personal data without instructions and law does not prohibit informing the controller on grounds of public interest, the processor shall inform the controller of that legal requirement before processing. (Art. 28(3)(a)) |
| R11 (PO5)      | The processor shall ensure that persons authorized to process personal data have committed themselves to confidentiality or are under an appropriate statutory obligation of confidentiality. (Art. 28(3)(b)) |
| R12 (PO6)      | The processor shall take all measures required pursuant to Article 32 or to ensure the security of processing. (Art. 28(3)(c)) |
| R13 (PO7)      | The processor shall assist the controller in fulfilling its obligation to respond to requests for exercising the data subject’s rights. (Art. 28(3)(e)) |
| R14 (PO8)      | The processor shall assist the controller in ensuring the security of processing. (Art. 28(3)(f), Art. 32) |
| R15 (PO9)      | The processor shall assist the controller in notifying a personal data breach to the supervisory authority. (Art. 28(3)(f), Art. 33) |
| R16 (PO10)     | The processor shall assist the controller in communicating a personal data breach to the data subject. (Art. 28(3)(f), Art. 34) |
| R17 (PO11)     | The processor shall assist the controller in ensuring compliance with the obligations pursuant to data protection impact assessment (DPIA). (Art. 28(3)(f), Art. 35) |
| R18 (PO12)     | The processor shall assist the controller in consulting the supervisory authorities prior to processing where the processing would result in a high risk in the absence of measures taken by the controller to mitigate the risk. (Art. 28(3)(f), Art. 36) |
| R19 (PO13)     | The processor shall return or delete all personal data to the controller after the end of the provision of services relating to processing. (Art. 28(3)(g)) |
| R20 (PO14)     | The processor shall immediately inform the controller if an instruction infringes the GDPR or other data protection provisions. (Art. 28(3)(h)) |
| R21 (PO15)     | The processor shall make available to the controller information necessary to demonstrate compliance with the obligations Article 28 in GDPR. (Art. 28(3)(h)) |
| R22 (PO16)     | The processor shall allow for and contribute to audits, including inspections, conducted by the controller or another auditor mandated by the controller. (Art. 28(3)(h)) |
| R23 (PO17)     | The processor shall impose the same obligations on the engaged sub-processors by way of contract or other legal act under Union or Member State law. (Art. 28(4)) |
| R24 (PO18)     | The processor shall remain fully liable to the controller for the performance of sub-processor’s obligations. (Art. 28(4)) |
| R25 (PO19)     | When assessing the level of security, the processor shall take into account the risk of accidental or unlawful destruction, loss, alteration, unauthorized disclosure of or access to the personal data transmitted, stored or processed. (Art. 32(2)) |
| R26 (CR1)      | In case of general written authorization, the controller shall have the right to object to changes concerning the addition or replacement of sub-processors, after having been informed of such intended changes by the processor. (Art. 8(2)) |

\(^1\) Cat is the category of the requirement, namely metadata (MD), processor’s obligation (PO), controller’s right (CR), and controller’s obligation (CO).

\(^2\) GDPR-related articles.

In collaboration with legal experts from Linklaters, we extracted from GDPR a total of 45 compliance requirements that regulate DPAs. A GDPR-compliant DPA provides assurance about the completeness of the requirements concerning data processing activities, which are essential for developing compliant software systems. According to the experts’ feedback, there are 26 mandatory and 19 optional requirements. The different requirements’ types were discussed and agreed upon during several sessions.

Mandatory requirements (R1 – R26, listed in Table 1) are explicitly stipulated in GDPR, except R1 and R2 which were strongly recommended by legal experts. R1 and R2 provide key information concerning the identity and contact details of the data controller and processor based on which the compliance checking is performed. Optional requirements (R27 – R45, see supplementary material) are derived from good practice in writing DPAs according to legal experts. Our automated support results in a non-compliance decision when a mandatory requirement is violated, and raises a warning when an optional requirement is violated.

We further group the compliance requirements into four categories, namely nine metadata requirements (denoted as...
MD), 25 requirements concerning the processor’s obligation (denoted as CO), and eight about the controller’s rights (denoted as CR). As shown in Table 1, the majority of mandatory requirements describe the processor’s obligations (corresponding to 19/25).

To build our artifacts, we followed an iterative method that includes three steps. The first step was reading the articles of GDPR related to DPA. In the second step, we created the artifacts. Finally, we validated these artifacts with legal experts. Through a close interaction with these experts, we incrementally improved and refined the artifacts to get them in their final form. To mitigate fatigue, we conducted each validation session for a duration of approximately two hours adding up to an overall of 46 hours for all sessions. Three legal experts participated in these joint sessions. Their expertise spans several areas including corporate laws, commercial litigation, data protection laws and financial regulations, and information technology within the legal domain. During these sessions, we refined the requirements according to the legal experts’ interpretation, thereby alleviating any ambiguity in the GDPR. During these steps, we created the glossary terms (the second artifact provided as supplementary material) with traceability links to GDPR articles to facilitate the understanding of the requirements.

Following experts’ recommendations, we started by analyzing Art. 28 in GDPR as it is directly relevant to the compliance of DPAs. We then read through all referenced articles in Art. 28, namely Articles 32–36, 40, 42, 43, 63, 82–84, and 93(2). While reading the articles, we (1) transformed the text in GDPR articles into “shall” requirements and (2) divided compound statements into simple requirements. For example, we extracted from Art. 28(3), which states that “Processing by a processor shall be governed by a contract [...] that sets out the subject-matter and duration of the processing, the nature and purpose of the processing, the type of personal data and categories of data subjects [...]”, four requirements related to the metadata content of a DPA, marked as R3 – R6 in Table 1. We adopted the GDPR text as-is when it resembled a “shall” requirement such as R34 (see supplementary material) extracted from Art. 33(2).

Using the resulting list of 45 compliance requirements, we manually checked the compliance of 24 DPAs. Specifically, we identified for each requirements whether it is satisfied by any statement in the DPA. As we elaborate in Section 5.2 our manual analysis aimed at establishing in-depth understanding of the compliance requirements before outsourcing the annotation to external annotators.

4 Automated Approach for DPA Compliance Checking against GDPR (RQ2)

To address RQ2, we devise an NLP-based approach for checking the compliance of DPAs against GDPR at a phrasal level (thereafter referred to as DERECHA, standing for DPA sEmantic fRamE-based Compliance cHecking Against GDPR). The current version of DERECHA does not use machine learning or deep learning architectures as, in our application context, we typically do not have enough labeled data to train robust models with accurate predictions. Recent advances in the NLP literature show that fine-tuning large-scale language models like BERT often yields accurate models for many downstream tasks. To this end, we note that such language models are integrated into currently available off-the-shelf semantic role labeling tools which DERECHA does not directly use as we explain later in this section. Instead, we employ these off-the-shelf NLP tools as a baseline solution and compare its performance in Section 5.3 against the performance of DERECHA. That said, we believe that attempting to extend our dataset and experimenting with machine learning or deep learning models would be important to refine our conclusions. This is left for future work.

Fig. 2 presents an overview of DERECHA, composed of five steps. The input to DERECHA includes a DPA, the name(s) of the controller and processor, and a set of requirements. The output is then a compliance report which summarizes the findings of DERECHA. In step A, we manually create SF-based representations of the compliance requirements extracted in Section 3. Note that this step is performed only once and can be reused across DPAs. In practice, from a user perspective, the SF-based representations identified in this paper can be adopted as-is. In step B, we apply an NLP pipeline to parse the DPA and preprocess the text. In step C, we automatically generate SF-based representations for the input DPA. In step D, we enrich the textual content in the DPA with semantically-related text, e.g., the text “engage” will be enriched with “hire” and “employ”. In the last step, step E, we use the SF-based representations created in steps A and D to check the compliance of the input DPA against GDPR. Finally, we generate a detailed report about the compliance decision. We elaborate these steps next.

4.1 Step A: Defining SFs in Compliance Requirements

In this step, we define an SF-based representation to characterize the information content of each requirement in GDPR. This representation aims at decomposing the requirement and thus enabling compliance checking at a phrasal level.

As can be seen in Table 1, the metadata requirements describe the content that a DPA has to cover, e.g., the duration of the agreement. In contrast, the requirements under the other categories (e.g., processor’s obligations) are expected to be precisely stated in the DPA. While the exact terminology is not required by GDPR, the semantics of these requirements are. This has an implication on the compliance checking process. For example, a statement in a given DPA satisfying R5 (MD5) will unlikely mention “the DPA” or “shall contain”. On the contrary, a statement satisfying R9 (PO3) has to explicitly cover the different elements of its SF-based representation illustrated in Fig. 2. Consequently, we define such representations only for the requirements concerning processor’s obligations, controller’s rights and controller’s obligations. We elaborate on how we handle the compliance checking against the metadata requirements (R1 – R6 and R27 – R29) in Section 4.5.

To define the SF-based representation, we first identify the SRs in each requirement. Inspired by existing literature, we collate a list of 10 SRs pertinent to DPA compliance requirements in GDPR. Four SRs are illustrated in Fig. 2, including action to describe a particular event, and
actor, object, and constraint to describe the SRs of the participants in that event. The remaining six SRs, exemplified on R18 – R20 in Table I are: Beneficiary – an entity that benefits from an action (e.g., [the controller]beneficiary in R18); reason – justifying why the action is performed (e.g., [to mitigate the risk]reason in R18); time – a temporal aspect (e.g., [prior to processing]time in R18); condition – the cases where the action is performed (e.g., [If an instruction infringes GDPR or other data protection provisions]condition in R20); situation – something that happened or can happen (e.g., [the end of provision of services related to processing]situation in R19); and finally reference – the mention of legal entities (e.g., [GDPR or other data protection provision]reference in R20).

Recall from Section 2.2 that an SF is represented as a predicate-arguments structure. For each requirement, we therefore define a predicate and a set of arguments evoked by that predicate. The predicate highlights the main action a data processor needs to perform to be compliant with GDPR. Example predicates include (not engage) in R7, (process) in R9, and (return or delete) in R19. The arguments are the different phrases in the requirement which constitute the participants associated with the predicate. We manually identify the arguments by applying the list of SRs explained above on the requirement text. Different requirements can share the same predicate when they refer to the same action, e.g., six requirements share the predicate (assist), namely R13 – R18. Nonetheless, we define one representation for each requirement, since the required arguments might be different. Doing so is essential for capturing missing information content in DPA statements according to what is expected by each requirement. Finally, we have discussed the SF-based representations with our collaborating experts from Linklaters and received feedback that the underlying SRs are intuitive and easy to understand. This suggests extending our approach with more compliance requirements alongside their respective SF-based representations is practically feasible.

The first step results in: (i) a list of SRs collated and refined from the existing literature which is passed on to step C to be further used in automatically generating SF-based representations for the input DPA; and (ii) SF-based representations for the requirements (R7 – R25 and R30 – R45) passed on to step E. These representations enable automated compliance verification of the statements at a phrasal level. An example of such representation for PO6 (R12 in Table I) contains the predicate (take), and a set of four arguments $A = \{a_1 : 0 \leq i \leq 3\}$ such that $a_0 = [\text{the processor}]actor, a_1 = [\text{all measures}]object, a_2 = [\text{Article 32}]reference, a_3 = [\text{to ensure the security of processing}]reason.$

The SF-based representations for all 45 requirements are provided as supplementary material.

4.2 Step B: Preprocessing

Given a DPA as input, this step applies an NLP pipeline composed of seven modules, including: (1) tokenization, splits the text into tokens; (2) sentence splitting, identifies the different sentences in the text using different heuristics, e.g., a sentence ends with a period; (3) lemmatization, maps the words to their canonical forms, e.g., “instructions” and “requested” are mapped to “instruction” and “request”; (4) POS tagging assigns a part-of-speech (POS) tag to each token, e.g., noun, verb, and possessive pronoun; (5) chunking, groups the words that form a syntactic unit, e.g., “the processor” is a noun phrase; (6) dependency parsing, describes the grammatical relations in a sentence, e.g., “the processor” and “personal data” in the sentence “the processor transfers personal data” are the subject and the object of the verb “transfers”, respectively; and finally (7) semantic parsing, defines the meanings of the different constituents in a sentence. For semantic parsing, we apply two widely-known lexical resources, namely WordNet [25, 26] and VerbNet [27, 28]. WordNet is a lexical database for English. It groups words into sets of synonyms called synsets and further connects the synsets via semantic relations. For example, is-a connects a hyponym (more specific synset) to a hypernym (more general synset) [29]. VerbNet groups verbs into classes according to their structure, e.g., the verbs under the motion class include run, escape, and leave [30].

Each module in this pipeline generates NLP annotations on parts of the text in the input DPA. The resulting NLP annotations are passed on to the next steps.

4.3 Step C: Generating SFs Automatically

In step C, we automatically generate for each statement in the input DPA an SF-based representation. To do so, we first identify the SRs in the DPA statement. Subsequently, we use the SR action to generate the predicate and the remaining SRs to generate the arguments. Recall from Section 2.2 that a statement in a DPA corresponds to a sentence.

In this step, we design our own method for identifying SRs instead of directly using the results of available NLP tools for semantic role labeling. First, the NLP tools have been trained on a large body of generic text. This can result in missing some of the SRs in our list which are pertinent to compliance checking against GDPR, e.g., reference and situation. Second, a typical NLP semantic role labeler would generate a predicate for each verb in a given statement.
In contrast, our representation is based on generating one predicate for each statement in DPA, i.e., extracting one SR action to which other SRs in the statement should be related. Generating one predicate is essential for the purpose of our analysis since the SF-based representations of GDPR requirements are centered around one predicate. Our goal in this step is to decompose each statement in the DPA into meaningful phrases that are labeled with similar SRs as in the requirements. In addition, comparing multiple predicates for each statement in the DPA against one predicate of each requirement is far too time consuming as reported in Section 5.3. Further, this would lead to multiple answers, most of them being irrelevant, as explained in Section 5.3.

To empirically assess the above limitations with respect to the performance of our approach, we define a baseline that involves two sub-steps, as explained below.

**C1) SRs identification.** To identify SRs in a given statement, we utilize the NLP annotations from step A (Section 4.2). In particular, we apply a set of rules (listed in Table 2) over the syntactic information of the statement. We rely mostly on the grammatical relations produced by the dependency parser. For instance, the action which will be regarded as the predicate is the main verb in the statement. Consequently, the actor should be associated with the action as its subject. We note that any statement without a root verb (e.g., the title of a section) will not have a predicate (i.e., action) and is thus not represented by our method.

Our rules incorporate a set of markers that help identify the SRs. Examples of markers are shown in Table 3. Except for situation and reference, the markers we use in our work originate from the RE literature [23], [24]. We improved these markers in two ways. First, we included the nouns and/or verbs that occur in the respective SRs according to our SF-based representations defined for GDPR requirements in step A. The intuition is to cover what is expected by the GDPR requirements for compliance. Second, we enrich the resulting markers (including the ones reported in the literature and the ones we adapted from GDPR) by including their synonyms from WordNet and VerbNet. This way, our markers will likely cover the different wordings applied in the DPA text.

**C2) Text spans demarcation.** Once an SR is identified, we demarcate the text span to which this SR should be assigned. Specifically, we use the NLP annotations produced by the text chunking to find the entire phrase where the SR is located. For instance, the SR action is assigned to the verb “employ” in the statement displayed in Table 4. We then find the verb phrase in the chunking results which contains “employ”, i.e., “can employ”. This way, we ensure capturing the semantics of the statement in a proper way. Except for action which has to be identified in a statement only once, other SRs can be identified in the statement multiple times. In this case, we group the n text spans under that SR to get the argument $[span_1 | \ldots | span_n]_{role}$. For example, the argument $[after | end]_{time}$ contains two spans related to time.

As a result of the sub-steps C1 and C2 explained above, each statement in the input DPA is sliced into a set of phrases, each one with an SR label. These phrases constitute our SF-based representation for the statement which is further processed in the next step.

### 4.4 Step D: Enriching Text of Input DPA

In this step, we enrich the DPA text by extending the text spans demarcated in the previous step with semantically-related words. The rationale is to increase the likelihood of finding an overlap when comparing the DPA text against GDPR requirements. We retrieve semantically-related words from the annotations of the semantic parsing module in Section 4.2. We extract from WordNet synonyms for each word in a text span labeled with any SR in the statement. For example, an object spanning “intended changes” will be enriched with the related words “planned, alteration, modification”. To enrich the text span of action, in addition to WordNet, we use the verbs provided in VerbNet over the same class, e.g., the action spanning “can employ” will be enriched from VerbNet with the verbs “engage” and “hire”, since they are grouped under the same class. To select the synonyms of the words in this step, we opted for the most frequent sense (MFS) of the word. While disambiguating the words can be beneficial, it is outside the scope of our paper. We note that MFS has been widely used as a baseline in the NLP literature and has often shown sufficient performance [51].

In this step, we further replace the actual names of the controller and processor with the generic place holders “the controller” and “the processor”, e.g., *Levico Accounting GmbH* in the example in Fig. 1 is replaced with “the processor”. As explained in Section 1, we assume in our work that the name(s) of the controller and processor are given as input by the user. Replacing those named entities helps normalize the text and improve the overall compliance checking.

### 4.5 Step E: Compliance Checking

In step E, we check the compliance of the input DPA against GDPR.

**Metadata Requirements.** To verify whether the DPA satisfies the metadata requirements, we do the following. R1 and R2 are concerned with the identity and contact details of the controller and processor. Given the names as input by the user, we look for statements in the DPA that contain these names. To find the contact details, we create regular expressions that identify entities including phone numbers, email and postal addresses. Then, we check whether these entities are present in the same statements containing the names of the controller and processor.

For verifying the remaining requirements (R3 – R6 and R27 – R29), we apply a method inspired by Lesk algorithm [32], one of the traditional methods applied in NLP in the context of word sense disambiguation. The original Lesk algorithm compares multiple senses (meanings) of two words for the purpose of identifying their respective meanings. For example, to disambiguate the word “cone” in the phrase “the pine cones”, Lesk algorithm compares the definition of each sense of the word “pine” with the definition of each sense of the word “cone”. The algorithm
TABLE 2: Extraction Rules of Semantic Roles.

| SR   | Description (D) and Example (E) from Fig. [4] |
|------|-----------------------------------------------|
| Action | (D) The action is the VP that contains the root verb retrieved from the dependency parsing tree. The action is the main verb in a statement. (E) The action in S6 is “processes”. |
| Actor  | (D) The actor is the NP containing the subject of the root verb in the statement (i.e., the action). (E) The actor in S6 is “Levico Accounting GmbH” |
| Object | (D) The object is either the NP that contains the object of the root verb AND the action does not contain any beneficiary marker; OR the VP that starts with a preposition and is associated with the root verb. Our rule accounts for cases where the root verb is directly followed by a PP (e.g., “assist with”). To differentiate actions that evoke the object (e.g., “process”) from other actions that evoke beneficiary (e.g., “inform”), we define markers. (E) The object in S6 is “Company personal data” |
| Beneficiary | (D) Similar to object, the beneficiary is either the NP that contains a preposition linked to the root verb AND the action contains a beneficiary marker, OR the NP that contains a subject of a verb that is different from the root verb. Since the beneficiary is an entity that benefits from the action, it can be associated directly with the action given that the action has a marker that pinpoints the necessity of a beneficiary. Alternatively, the beneficiary can be an entity associated with another action in the statement. (E) The beneficiary in S11 is “Company” |
| Condition | (D) The condition is the PP, OR ADVP, OR any phrase annotated as subordinating clause by the dependency parser given that it contains a condition marker. (E) The condition in S11 is “without undue delay” |
| Constraint | (D) The constraint is identified in similar manner as condition except that the phrase should contain a constraint marker. The constraint in S11 is “without undue delay” |
| Time | (D) The time is the NP, OR PP, OR ADVP, OR any phrase annotated as conjunctive preposition by the dependency parser given that it contains a time marker. (E) The time in S12 is “remains effective so long as Levico provides services” |
| Reason | (D) The reason is either the VP containing an auxiliary and is linked to a verb other than the root, OR the VP containing an open clausal complement and is linked to the root verb. (E) The reason in S6 is “for providing the service of payroll administration” |
| Situation | (D) The situation is the NP OR VP that contains a situation marker. (E) The situation in S13 is “termination of any services” |
| Reference | (D) The reference is the NP that contains a reference marker. (E) The reference in S10 is “Article 32(1)” and “GDPR”. |

† VP: verb phrase, NP: noun phrase, PP: prepositional phrase, ADVP: adverbial phrase.
‡ The extraction rules are adapted from the RE literature [23, 24].

TABLE 3: Example Markers for Identifying Semantic Roles.

| SR   | Markers |
|------|---------|
| Beneficiary | inform, report, assist, help, aid, support, remain, notify, provide, give, supply |
| Condition | if, once, in case, where, when |
| Constraint | without, on, in accordance with, according to, along, by, under, unless |
| Time | after, later, prior, before, earlier, as long as, as soon as |
| Situation | access, addition, destruction, loss, disclosure, adherence, modification, termination, expiration |
| Reference | gdpr, dpa, law, jurisprudence, legislation, agreement, article, contract |

then computes the number of overlapping words between the definitions of each two senses. The senses with the highest overlap are the ones selected to disambiguate the two words. We adapt this algorithm to our compliance checking process. Specifically, for each metadata requirement (among R3 – R6 and R27 – R29), we look for a statement in the DPA that has overlapping words with the requirement. During our manual analysis of the 24 DPAs (discussed in Section 3), we observed that DPA statements often apply the same words when expressing the requirements R3 – R6 and R27 – R29. Thus, we believe that using Lesk algorithm in this case is sufficient.

Remaining Requirements. To check whether a requirement is satisfied in the DPA, we compare the SF-based representation of the requirement against that of each statement in the DPA. Table 2 shows a simplified example of checking compliance between a DPA statement and requirement R7. Our compliance checking method using SFs comprises four steps, as elaborated in Algorithm 1. First, we check the predicates in the two representations. Only if the predicates are sufficiently similar, we subsequently match the arguments, and further compute a score indicating the matching degree. Finally, we conclude the compliance decision for the DPA. We explain each step next.

E.1) Matching predicates. For matching two predicates, we specifically consider the verbs occurring in these predicates. We then deem the predicates to be similar if at least one of the two conditions below is met (line 6 in Algorithm 1). Condition 1 returns true if the two predicates have some overlapping verbs.

Condition 2 returns true if the semantic similarity between the verb in the requirement and at least one verb in the statement is greater than or equal to a threshold $\theta_p$. In our work, we apply the similarity metric (WuP) proposed by Wu and Palmer [33] and widely-applied in NLP applications [34–36]. In brief, WuP is a WordNet-based metric that computes the similarity between two words (or concepts) by finding the path length from the first hypernym in the hierarchy that is shared between the synsets of these concepts [37]. WuP returns a score between 0.0 (dissimilar) and 1.0 (identical).

To optimize the efficiency of DERECHA, only statements having similar predicates to the requirements will be further checked for compliance. A statement that meets none of the conditions above is likely discussing an action that is irrelevant to the one required by GDPR.
TABLE 4: Example of Compliance Checking using SF-based Representations against R7.

| GDPR | | | | |
|---|---|---|---|---|
| actor | (p) | object | constraint | time |
| the processor | not engage | a sub-processor | without a prior | specific or general written authorization of the controller. |

Levico Accounting GmbH can employ sub-contractors for the performance of the service of Levico’s obligations.

DPA*

| actor | (p) | object | reason | constraint |
|---|---|---|---|---|
| Levico Accounting GmbH | can | hire | engage | MISSING |

E.2) Matching arguments. This is done based on the arguments in the compliance requirement, since they represent required information by GDPR (lines 7–18 in Algorithm 1). In particular, we align each argument in the requirement to an equivalent argument sharing the same SR in the statement (e.g., actor in Table 4).

Any argument whose SR exists in the statement but not in GDPR is considered as not required. Conversely, an argument is missing if its SR is present in GDPR but not in the statement. For example, the argument with the SR reason in Table 4 is not required by R7, whereas the one related to constraint (including time) is missing from the statement. For each pair of equivalent arguments, we check for overlap between the two text spans in a similar way to condition 1 and condition 2 above. For condition 2, we compute the Jaro-Winkler distance instead of WuP that we applied for matching predicates. The reason is that Jaro-Winkler distance is more suitable in the case of matching arguments since longer text sequences are expected unlike the case of predicates.

Finally, equivalent arguments are deemed to be matching if there is any overlap between the text spans or when the similarity between the text spans is greater than a threshold \( \theta_a \). The arguments are deemed not matching otherwise.

Following the findings, we empirically set \( \theta_p \) to 0.9 and \( \theta_a \) to 0.7. We note that for \( \theta_p \), unlike \( \theta_a \), we experimented with high values (> 0.7). The rationale is that arguments are only checked when the predicates are deemed similar, i.e., the similarity value between the two predicates passes the threshold \( \theta_p \). Considering lower threshold values for \( \theta_p \) would entail significant execution time for performing more comparisons involving arguments matching. To avoid excessive computations and further ensure a high level of similarity, we set the first threshold (\( \theta_p \)) to be relatively high. Once the similarity between two predicates exceeds the threshold (i.e., >0.9), we match the arguments. The second threshold (\( \theta_a \)) is set a bit lower since we already have evidence about the similarity of the predicates.

In the example shown in Table 4, the predicates are matching since both predicates contain the overlapping verb “engage”. For the sake of providing an example, though
not needed here as condition 1 is met, for condition 2, we compute WuP between the verb in the predicate of R7 and each verb in the enriched predicate of the statement. This results in WuP scores of 0.4 between “engage” and “hired”, 0.4 between “engage” and “employ”, and 1.0 between “engage” and “engage”. In a similar manner, we match the arguments of R7 with the ones in the statement. Given the overlap in the text spans of the actor and object, they are considered to be matching. The argument related to reason in the statement is skipped since it is not part of the requirement. The argument related to constraint has no equivalent argument in the statement and is thus considered not matching.

E.3) Computing matching degree. DERECHA further computes a score for each compliance requirement indicating the maximum matching degree of any statement satisfying that requirement in the DPA (line 23 in Algorithm 1). Any statement that does not satisfy any of the two predicate matching conditions above directly receives a score of 0: the statement does not match the requirement at all. Otherwise, if a statement satisfies one predicate matching condition, then the matching degree is computed as a fraction where the numerator equals the total number of arguments found to be matching between the statement and the requirement, and the denominator is the total number of arguments expected by the requirement. We add one to both the numerator and denominator corresponding to the matching predicate since the predicate is an essential element of the SF. Including the predicates when computing the matching degree accounts for the cases when a statement matches only the predicate but none of the arguments in the requirement. In this case the statement should still receive some low matching degree that is greater than zero. For example, R7 in Table 3 expects a total of four arguments, namely actor, object, constraint, and time. The matching degree between the statement and R7 is 3/5 (i.e., 0.6), since the statement successfully matches the predicate and two expected arguments of the requirement.

A requirement can be satisfied by multiple statements in a DPA. In this case, we assign to the requirement the maximum matching degree score across statements. Statements with low matching degrees are due to the lack of matching arguments. However, since these statements have similar predicates to a requirement, they can still be relevant. For such statements, we therefore consider the matching degree as an indicator of the confidence level at which DERECHA predicts that any requirement is satisfied or not in the DPA. This can guide the analyst in verifying and possibly correcting decisions when the confidence level is low. We elaborate on this in Section 5.4.

E.4) Compliance decision. Finally, we make the compliance decision at the DPA level. A requirement is marked as satisfied in the DPA if there is at least one statement satisfying the requirement. Otherwise, the requirement is marked as violated. DERECHA recommends that the DPA should be considered as not compliant when at least one mandatory requirement is violated. When optional requirements are violated, DERECHA raises a warning to draw the attention of the analyst to missing common practices. Fig. 4 shows the detailed report corresponding to the DPA example in Fig. 7 as generated by DERECHA.

5 Evaluation

This section presents the empirical evaluation for our compliance checking approach.

5.1 Implementation

We have implemented our approach in Java. We extract the textual content of a DPA (provided as MS Word document) using Apache POI 3.17 [41]. For operationalizing the NLP pipeline applied in step B of Fig. 3, we use the DKPro 1.10 toolkit [42]. In particular, our operationalization employs the OpenNLP tokenizer, POS-tagger and text chunker [43], Mate lemmatizer [44] and Stanford dependency parser [45]. For enriching the textual content of DPA statements (step D), we rely on the semantic parsing results produced by the NLP pipeline. In particular, we apply extJWNL 1.2.46 for accessing WordNet, and JVerbNet 1.2 [47] for accessing VerbNet. To compute semantic similarity needed in step E in our approach, we use the Jaro-Winkler distance [48] and the wup metric [33] as implemented by the WS4J 1.0 library [49].

5.2 Data Collection Procedure

To conduct this study, we have collected a total of 54 DPAs covering diverse sectors and services such as telecommunication, banking, healthcare, postal services, data analytics, cloud and web hosting services. Among them, 34 DPAs have been provided by our industry collaborator (Linklaters LLP) and the rest have been collected from online resources. The goal of our data collection is to manually verify whether
a given DPA complies with GDPR requirements and more specifically what statements satisfy which requirements. Our data collection was performed in two phases described next.

In the first phase, three authors of this paper manually analyzed 24 DPAs for compliance with GDPR. This analysis spanned two working weeks. We started by independently analyzing five DPAs. Specifically, we checked each sentence in the DPAs against the 45 compliance requirements defined in Section 3 and further labeled the sentence with the requirement(s) it satisfied. We then met and discussed our findings and observations in several joint sessions. Once we reached an agreement, we split the remaining DPAs such that each DPA was analyzed by two authors. Finally, we carefully discussed with our collaborating experts the statements that were not straightforward. Our interactions with the legal experts were divided into four sessions over the course of one month, where each session lasted two hours. During this phase, we acquired knowledge about the compliance checking of DPAs and devised annotation guidelines for the second phase.

The second phase had three third-party annotators read through and examine the remaining 30 DPAs for compliance against GDPR. All of the annotators are law students and went through a half-day training on GDPR compliance. We shared with the annotators the guidelines from the first phase, the DPAs to be analyzed and the list of compliance requirements. The annotators were instructed to examine each statement in the DPA and select all requirements which it satisfied. We then met and discussed our annotations over a one-month span, during which they declared an average of 40 working hours. To mitigate fatigue, the annotators were recommended to limit their work to a maximum of two hours per day.

To measure the quality of the annotations in this phase, we computed the interrater agreement using Cohen’s Kappa (κ) [50] on five of the 30 DPAs. The average κ value across pair-wise agreements of the annotations, split by categories metadata, processor’s obligations, controller’s obligations, and controller’s rights, are 0.77, 0.72, 0.82, and 0.79, respectively. κ value for the category controller’s obligations is “almost perfect agreement” [51], whereas the rest of the three categories indicate “moderate agreement” among the annotators. The disagreements were discussed and resolved by the annotators in a subsequent session.

The overall document collection resulted in analyzing a total of 7048 statements in both phases, out of which 1742 (25%) are marked as satisfying at least one requirement. The DPAs annotated during the first phase were the basis for developing our approach (thereafter referred to as Dev dataset), whereas we used the DPAs annotated by the third-party annotators (thereafter referred to as EV) for addressing our research questions (RQs) and evaluating the performance our approach. Table 5 shows the results of our document collection considering the entire dataset, Dev and EV. The compliance requirements are sorted by category (with mandatory requirements in boldface) and, for each requirement, the table reports the total number of DPAs (D) in which the requirement was found to be satisfied across the 54 DPAs and the total number of statements (S) that satisfy the requirement. For example, R9 describes the processor’s obligation (PO3) concerning processing personal data only on documented instructions from the controller. In our dataset, R9 was satisfied in 43 DPAs (out of 54 DPAs), and a total of 83 statements were used across the DPAs to comply with R9.

We observe from the table that mandatory requirements are often satisfied by DPAs while the optional ones are not. According to our collaborating experts, requirements concerning controller’s obligations are not necessary for compliance, since GDPR provisions related to DPA focus mostly on the metadata requirements and processor’s obligations. We note that, as stated earlier, despite our extensive data collection in terms of DPAs, Table 5 confirms that the small number of statements we have, for each requirement, prevents us from developing DERECHA based on machine learning.

| Cat† | ID | Total | Dev Data | Test Data (EV) |
|------|----|-------|----------|----------------|
|      | D  | S     | D        | S          |
| MD1  | R1 | 22    | 7        | 15         | 19          |
| MD2  | R2 | 40    | 13       | 50          | 27          | 43          |
| MD3  | R3 | 33    | 45       | 15          | 22          | 18          | 23          |
| MD4  | R4 | 38    | 133      | 11          | 58          | 27          | 75          |
| MD5  | R5 | 38    | 270      | 11          | 118         | 27          | 152         |
| MD6  | R6 | 36    | 126      | 10          | 31          | 26          | 95          |
| MD7  | R7 | 31    | 914      | 9           | 391         | 22          | 523         |
| MD8  | R8 | 21    | 99       | 5           | 22          | 16          | 77          |
| MD9  | R9 | 8     | 9        | 5           | 5           | 3           | 4           |
| PO1  | R7 | 34    | 62       | 16          | 24          | 18          | 38          |
| PO2  | R8 | 34    | 52       | 17          | 31          | 17          | 21          |
| PO3  | R9 | 43    | 83       | 14          | 30          | 29          | 53          |
| PO4  | R10| 27    | 55       | 12          | 24          | 15          | 31          |
| PO5  | R11| 44    | 64       | 18          | 29          | 26          | 35          |
| PO6  | R12| 36    | 206      | 14          | 49          | 22          | 157         |
| PO7  | R13| 37    | 85       | 14          | 50          | 23          | 35          |
| PO8  | R14| 15    | 22       | 12          | 17          | 3           | 5           |
| PO9  | R15| 18    | 25       | 11          | 14          | 3           | 3           |
| PO10 | R16| 20    | 27       | 15          | 16          | 7           | 9           |
| PO11 | R17| 32    | 34       | 11          | 22          | 5           | 5           |
| PO12 | R18| 15    | 17       | 12          | 12          | 21          | 22          |
| PO13 | R19| 39    | 67       | 13          | 28          | 26          | 39          |
| PO14 | R20| 23    | 25       | 8           | 9           | 15          | 16          |
| PO15 | R21| 29    | 48       | 17          | 29          | 12          | 19          |
| PO16 | R22| 35    | 49       | 14          | 15          | 21          | 34          |
| PO17 | R23| 38    | 59       | 16          | 29          | 22          | 30          |
| PO18 | R24| 31    | 53       | 14          | 27          | 17          | 26          |
| PO19 | R25| 28    | 36       | 13          | 17          | 15          | 19          |
| PO20 | R30| 26    | 29       | 14          | 17          | 12          | 12          |
| PO21 | R31| 15    | 19       | 7           | 8           | 8           | 11          |
| PO22 | R32| 24    | 35       | 12          | 21          | 12          | 14          |
| PO23 | R33| 17    | 23       | 9           | 15          | 8           | 8           |
| PO24 | R34| 37    | 67       | 10          | 25          | 27          | 42          |
| PO25 | R35| 18    | 25       | 6           | 8           | 12          | 17          |
| CR1  | R26| 24    | 34       | 13          | 22          | 11          | 12          |
| CR2  | R44| 4     | 5        | 3           | 3           | 1           | 2           |
| CR3  | R45| 8     | 21       | 5           | 16          | 3           | 5           |
| CO1  | R36| 1     | 1        | 0           | 0           | 1           | 1           |
| CO2  | R37| 2     | 3        | 2           | 3           | 0           | 0           |
| CO3  | R38| 1     | 1        | 0           | 0           | 1           | 1           |
| CO4  | R39| 1     | 1        | 0           | 0           | 1           | 1           |
| CO5  | R40| 1     | 1        | 1           | 1           | 0           | 0           |
| CO6  | R41| 1     | 1        | 1           | 1           | 0           | 0           |
| CO7  | R42| 1     | 1        | 0           | 0           | 1           | 1           |
| CO8  | R43| 9     | 14       | 5           | 7           | 4           | 7           |

† Cat: Requirement’s category; mandatory requirements are in bold.
5.3 Evaluation Procedure

To answer RQ3 and RQ4, we conduct the experiments explained below.

**EXPI.** This experiment addresses RQ3. EXPI evaluates whether our approach accurately identifies the violations of compliance requirements in a given DPA. Specifically, we compare the results per requirement in each DPA against the manual annotations in EV. We define for each requirement \( R_i \) a true positive (TP) when \( R_i \) is violated in the DPA and correctly marked as violated by DERECHA, a false positive (FP) when \( R_i \) is satisfied (i.e., not violated) in the DPA, but falsely marked \( R_i \) as violated, a false negative (FN) when \( R_i \) is violated in the DPA but falsely marked as satisfied, and finally a true negative (TN) when \( R_i \) is satisfied in the DPA and correctly marked as such. Following this, we report the accuracy \( A \), precision \( P \), and recall \( R \), computed as \( A = (TP+TN)/(TP+FP+FN+TN) \), \( P = TP/(TP+FP) \), and \( R = TP/(TP+FN) \), respectively. In addition, we report the F-score \( F \) computed as \((1+\beta^2)*(P*R)/(\beta^2*P+R)\). In EXPI, we choose to report \( F_{0.5} \), where \( \beta = \frac{1}{2} \), indicating that precision is more important than recall in the context of our study as we explain later in this section.

**Baseline.** In EXPI, we further compare our approach against a baseline (referred to as B1), which utilizes off-the-shelf NLP tools for producing SF-Based representations for both the compliance requirements and the DPA statements. More specifically, B1 applies existing NLP tools for extracting the SRs and generating the SF-based representations. The definitions of the SF-based representations for compliance requirements are not applicable for the baseline. The reason is that checking compliance requires aligning SF-based representations which are generated over the same set of SRs.

In B1, we apply the semantic role labeling module provided by AllenNLP [19], which is based on the BERT language model [19]. Since NLP tools generate a predicate-argument structure for each verb in the sentence, B1 generates automatically multiple SF-based representations for each statement in the DPA. B1 automatically generates as well the corresponding SF-based representations for each compliance requirement instead of applying our predefined representations. As a result, B1 checks the DPA compliance by comparing all SF-based representations generated by AllenNLP for a statement against those generated for a compliance requirement. For example, for the requirement “The processor shall not engage a sub-processor without a prior specific or general written authorization of the controller.” (R7 in Table 1), B1 generates two SF-based representations. The first one contains \{engage, \( A = \{\text{the processor}\_\text{argument0}, \text{a sub-processor}\_\text{argument1}\}\) and the second one is \{written, \( A = \{\text{authorization}\_\text{argument}\}\). B1 further enriches the DPA text similar to DERECHA. However, B1 limits matching predicates and aligned arguments to exact overlapping text due to efficiency concerns. Finally, B1 computes the proportion of predicates in a statement that are matching the predicates of the requirement. If this proportion is greater than a certain threshold \( \theta_{B1} \), then B1 concludes that the statement satisfies the requirement. In our experiments, we empirically tuned \( \theta_{B1} \) to 0.30. For the reasons explained in Section 4.2 SF-based representation does not apply for metadata requirements. Thus, B1 is applied only for the three other requirements categories.

**EXPII.** This experiment addresses RQ4. We run our approach on a laptop with a 2.3 GHz CPU and 32GB of memory. Our goal is to assess whether execution times suggest that DERECHA is applicable in practice.

5.4 Results and Discussion

**RQ3: How accurately can we check the compliance of a given DPA?**

Table[1] shows the results of EXPI, on the left-hand side, the results of DERECHA, and on the right-hand side the results of B1 introduced in Section 5.2. As explained in Section 5.2, the results are obtained by running DERECHA and B1 on the evaluation set (EV), which is comprised of 30 DPAs (7048 statements) on which we verify the compliance of our 45 requirements. DERECHA correctly finds 618 requirement violations (out of 750) and 524 satisfied requirements, but also leads to 76 false violations. As pinpointed in section 5.3, B1 is applied only to the categories of requirements PO, CO, and CR. Hence, we compare the performance of B1 against DERECHA considering only these categories. Specifically, DERECHA correctly identifies 550 violated requirements (out of 661) and 348 satisfied requirements (out of 419), while introducing 71 false violations. In comparison, B1 finds 436 violations, 243 satisfied requirements and introduces 175 false violations. To summarize, 182 errors (FVs+FNs) are produced by DERECHA for both mandatory and optional requirements in these three categories, whereas B1 produces 401 (219 more errors). Overall, DERECHA outperforms B1 by an average of \( \approx 20 \) percentage points (pp) in accuracy, \( \approx 17 \) pp in precision, recall, and \( F_{0.5} \).

In the context of our study, we favor high precision over high recall. In the case of unsatisfactory recall, the human analyst needs to go through the subset of statements that are marked by DERECHA as satisfying a requirement in order to determine whether this is actually the case. Unsatisfactory precision, on the other hand, indicates that some predicted violations cannot be trusted with a high level of confidence. In such a case, the human analyst would have no choice but to check all statements in the DPA to decide whether at least one of them actually satisfies a requirement. However, doing such a thing would defeat the very purpose of DERECHA. In contrast, addressing unsatisfactory recall can be done with relative ease as the number of statements satisfying any requirement usually represent a small percentage. We discuss next the trade-off between additional manual validation and accuracy improvement.

FNs entail that our approach can mispredict a requirement as being satisfied by at least one statement in the DPA. We carefully checked the matching degrees (scores defined in Section 4.5) of these FNs and observed that, in 86 out of 132 cases, such requirements are assigned matching degree scores \( \leq 0.5 \). By reviewing a fraction of statements, which are predicted to satisfy any requirement with a score below some predefined threshold, the analyst can identify statements that actually do not satisfy any requirement. One possible heuristic that was observed to work well is, for each requirement, to only analyze the statement with maximum score if it is below the selected threshold. Assuming we
TABLE 6: Results of Compliance Checking.

| ID   | Cat† | TPs | FPs | FNs | TNs | A   | P   | R   | F0.5 |
|------|------|-----|-----|-----|-----|-----|-----|-----|------|
| R1   | MD1  | 15  | 3   | 0   | 12  | 90.0| 83.3| 100 | 86.2 |
| R2   | MD2  | 3   | 0   | 0   | 27  | 100| 100 | 100 | 100  |
| R3   | MD3  | 5   | 0   | 7   | 18  | 76.7| 40.7| 78.1|
| R4   | MD4  | 1   | 0   | 2   | 26  | 93.3| 100 | 71.4|
| R5   | MD5  | 2   | 0   | 1   | 27  | 96.7| 100 | 66.7| 90.9 |
| R6   | MD6  | 2   | 0   | 2   | 26  | 93.3| 100 | 83.3|
| R7   | PO1  | 4   | 1   | 8   | 17  | 70.0| 80.0| 62.5| 5   |
| R8   | PO2  | 8   | 1   | 5   | 16  | 80.0| 88.0| 41.6| 12  |
| R9   | PO3  | 1   | 1   | 0   | 28  | 96.7| 50.0| 55.6| 0   |
| R10  | PO4  | 5   | 2   | 10  | 13  | 60.0| 71.4| 33.3| 58.1|
| R11  | PO5  | 4   | 1   | 0   | 25  | 96.7| 80.0| 100 | 83.3|
| R12  | PO6  | 2   | 0   | 6   | 22  | 80.0| 100 | 25.0| 62.5|
| R13  | PO7  | 3   | 1   | 4   | 22  | 83.7| 75.0| 42.9| 65.2|
| R14  | PO8  | 27  | 3   | 6   | 18  | 70.0| 73.9| 72.0| 72.3|
| R15  | PO9  | 17  | 3   | 6   | 4   | 70.0| 85.0| 73.9| 82.5|
| R16  | PO10 | 18  | 3   | 7   | 2   | 66.7| 85.7| 72.0| 82.6|
| R17  | PO11 | 3   | 3   | 6   | 18  | 70.0| 50.0| 33.3| 54.5|
| R18  | PO12 | 27  | 2   | 0   | 1   | 93.3| 93.1| 100 | 94.4|
| R19  | PO13 | 2   | 1   | 2   | 25  | 90.0| 66.7| 50.0| 62.5|
| R20  | PO14 | 12  | 5   | 3   | 10  | 73.3| 70.6| 80.0| 72.3|
| R21  | PO15 | 13  | 3   | 5   | 9   | 73.3| 81.3| 72.2| 79.3|
| R22  | PO16 | 4   | 1   | 5   | 20  | 80.0| 80.0| 44.4| 69.0|
| R23  | PO17 | 4   | 3   | 4   | 19  | 76.7| 57.1| 50.0| 55.6|
| R24  | PO18 | 13  | 3   | 0   | 14  | 90.0| 81.3| 100 | 84.4|
| R25  | PO19 | 10  | 4   | 5   | 11  | 70.0| 71.4| 66.7| 70.4|
| R26  | CR1  | 16  | 3   | 6   | 5   | 70.0| 84.2| 72.7| 81.6|

| ≈ Der  | CR2  | 24  | 3   | 3   | 0   | 80.0| 88.9| 88.9| 88.9|

| Summary | 618 | 76 | 132 | 524 | 84.6 | 89.1 | 82.4 | 87.6 | 436 | 175 | 226 | 243 | 62.9 | 71.4 | 65.9 | 70.2 |

† Cat: Requirement’s category; mandatory requirements are in bold.

apply the above heuristic, Fig. 5 shows the gain in recall achieved, for instance, when considering a threshold of 0.5: recall can be improved from 82.4% to 93.9% at the cost of reviewing ≈6% of the statements. For an average-sized DPA with 200 statements, this percentage corresponds to approximately 12 statements. Further increases in recall are possible but are associated with higher reviewing effort and the right trade-off depends on context. Eliminating FNs through manual reviews also slightly improves precision and accuracy. For example, the manual review mentioned above increases precision to 90.3% and accuracy to 91.0%. Selecting a threshold at which the user reviews DERECHA’s output (matching statements) in order to improve accuracy depends on the resources available and therefore the number of statements that can realistically be reviewed using our heuristic. Users who can afford it might opt for reviewing all of the matching statements with maximum scores for each requirement to achieve a near perfect recall (with ≈90.6% precision and ≈93.1% accuracy). In our DPAs, doing so would result in reviewing about 9% of statements on average across DPAs.

DERECHA achieves an average precision of 81.5% and 93.9% when identifying violations of mandatory (R1 – R26) and optional requirements (R27 – R45), respectively. In other words, DERECHA introduces false violations (FPs) in a total of 76 cases, out of which 50 are related to mandatory requirements. We analyzed these cases to determine the root causes, which we explain below. We further illustrate each cause with an example from our data collection, and report the number of FPs attributable to that cause.
Recall (%)  
84  86  88  90  92  94  96  99
Scores Threshold
0  1  2  3  4  5  6  7  8  9
Sentences (%)  
~6%  0.2  0.4  0.6  0.8  1.0  1.2  1.4  1.6  1.8
Fig. 5: Validation-accuracy trade-off.

C1. Inconsistent reference to the controller and/or processor: 22 FPs are caused by this error. We observed that in addition to using a concrete named entity representing the controller or the processor (e.g., Levico in Fig. 1), many DPAs use other words to refer to the controller or processor. For example, the following statement refers to the processor as data importer and to the controller as data exporter: “The data importer agrees at the request of the data exporter to submit its data processing facilities for inspection of the processing activities covered by the clauses which shall be carried out by the data exporter”. This statement should be predicted as satisfying R22 (see Table 1). However, it is unlikely for DERECHA to identify this statement correctly due to the way it is expressed and the additional unconventional names used in place of processor and controller. As an attempt to reduce such errors, we can modify DERECHA in a way such that the user of our tool can input multiple names for the controller or the processor accounting for both the named entity (e.g., Levico GmbH) and references (e.g., service provider). The user might be familiar with some commonly used names, e.g., contractor, client, customer, or operator.

C2. Loss of context: 39 FPs are caused by this error. In our work, we use the NLP pipeline to automatically demarcate statements in a given DPA, where each statement corresponds to one sentence as produced by the sentence splitter. Statements included in itemized or enumerated lists can suffer from loss of context when the statements cover grammatically incomplete sentences (marked as S1 – S5). For example, the following list spans five statements “Where Processor uses subprocessors for the processing of personal data on behalf of Controller, Processor shall: [S1]
- impose the same obligations on the engaged subprocessors; [S2]
- transfer only the data that is reasonably necessary for the purposes of providing the services; [S3]
- ensure all the time the security of the shared personal data, using encryption strategies; [S5]
- remain fully liable to the Controller for the performance of subprocessors.” [S5]
In the above example, the combination of the first two statements (S1 and S2) should be predicted as satisfying R23. However, considering these two statements separately leads to falsely predicting a violation of R23 instead.

C3. Noisy text in DPA: 15 FPs are caused by this error. Statements in a given DPA are often verbose, i.e., containing content that is considered irrelevant regarding compliance with requirements. Such cases can result in poor performance in the NLP syntactic parsing on which our SR extraction rules depend. For example, the following statement is long and contains multiple clauses that are hard to accurately parse: “Controller will permit Processor to take reasonable steps, according to a reasonable notice and during normal business hours, at Processor’s cost to assess compliance by Controller with its obligations under this DPA, including by inspecting Controller’s data processing facilities, procedures and documentation (limited to a maximum of one inspection in any twelve month period, or such further occasions as may be required by Privacy Law or if there is an actual Personal Data Breach by Controller)”.

Handling errors caused by C2 and C3 could be achieved with heuristics to identify and paraphrase itemized lists, or filter out noisy content from DPA statements.

Table 6 further suggests that the baseline B1 has two disadvantages. First, compared to our approach, B1 checks the compliance of a statement in DPA against a compliance requirement considering multiple verbs (predicates). This not only increases the processing time exponentially but also leads to poor results. Some of these predicates include very common verbs that can be found often in statements such as “mean”, “take”, and “write”. Having multiple predicates increases the likelihood of concluding a match between statements and requirements, consequently leading to more FNs (e.g., R14 and R33). Second, B1 relies fully on the NLP tool for semantic role labeling (SRL). In some cases, B1 could not identify any statement satisfying a given requirement since the SRL tool could not accurately parse the statement. This led to producing more FPs (e.g. R11 and R13). In contrast, DERECHA employs a set of rules that are likely to decompose the statement into meaningful phrases and thus conclude more accurately about their compliance against requirements. Our analysis shows that off-the-shelf NLP tools are unable to identify SRs in a legal text and are thus not adequate for checking the compliance of DPAs.

The answer to RQ3 is that DERECHA is 84.6% accurate in identifying both satisfied and violated GDPR requirements in DPAs. Specifically, DERECHA identifies violated requirements with a precision of 89.1% and a recall of 82.4%. Compared to a baseline that relies on existing NLP tools, DERECHA provides an accuracy gain of ≈20 pp. Our approach also computes matching scores that allow the analyst to review a small percentage of statements with low scores and thus eases the identification of FNs to further increase accuracy.

RQ4: How efficient is our approach in terms of the execution time?

Recall that our approach consists of five steps as shown in Fig. 3. Step A is concerned with manually defining SF-based representations for the requirements in GDPR. Step A took about a month, including reviewing related work to
define the set of SRs.

This step is a one-off and performed only occasionally upon changes in the regulation, i.e., new requirements become relevant to DPA compliance. For analyzing the complete evaluation set of 30 DPAs, running our approach requires \( \approx 90 \) minutes (corresponding to an average of \( \approx 0.77 \) second for processing a given statement). The detailed execution time required by each step of our approach is as follows. Preprocessing (step B) requires a total of 1 minute (\( \approx 0.009 \) second per statement). Generating the SF-based representations automatically (step C) is dominated by identifying SRs. This step takes in total about 52 minutes (\( \approx 0.44 \) second per statement). We note that step C consumes most of the time in our approach since our rules are built over syntax parsing. Finally, steps D and E require 37 minutes (\( \approx 0.32 \) second per statement). For efficiency, we implement step D (text enrichment) as part of step E (compliance checking), i.e., we perform text enrichment of arguments upon changes in the regulation, i.e., new requirements become relevant to DPA compliance. For analyzing the complete evaluation set of 30 DPAs, running our approach requires \( \approx 90 \) minutes (corresponding to an average of \( \approx 0.77 \) second for processing a given statement). The detailed execution time required by each step of our approach is as follows. Preprocessing (step B) requires a total of 1 minute (\( \approx 0.009 \) second per statement). Generating the SF-based representations automatically (step C) is dominated by identifying SRs. This step takes in total about 52 minutes (\( \approx 0.44 \) second per statement). We note that step C consumes most of the time in our approach since our rules are built over syntax parsing. Finally, steps D and E require 37 minutes (\( \approx 0.32 \) second per statement). For efficiency, we implement step D (text enrichment) as part of step E (compliance checking), i.e., we perform text enrichment of arguments only when the statements in DPA have passed predicate matching. Our approach runs \( \approx 10 \) times faster than the baseline \( B_1 \). The reason is, \( B_1 \) compares all the predicates in a given statement against all the predicates in a requirement leading to a significantly longer computational time.

To conclude, the answer to RQ4 is that our approach has a running time enabling its application in practice. Running our approach on an average-sized DPA (with 200 statements) takes \( \approx 2.5 \) minutes to verify its compliance against all 45 requirements.

6 RELATED WORK

Regulatory compliance is a long-standing research topic in the RE literature \([52]–[59]\). While little attention has been given to the compliance of DPAs, recent research explores GDPR compliance of data sharing. We therefore discuss current research on data sharing platforms (e.g., cloud service providers) since a GDPR-compliant DPA provides a complete set of compliance requirements that must be addressed in data processing activities. In addition, we address below two other broad topics that are closely related to our proposed approach for checking compliance, namely elicitation of compliance requirements from regulation, and automated support for compliance checking of requirements and software systems. Naturally, we focus our discussions in this section on GDPR.

6.1 GDPR Compliance of Data Processing Activities

Several studies analyze the impact of GDPR on data sharing platforms \([60], [61]\). Urban et al. \([62]\) study the behavior of online advertising companies in data tracking and information sharing with respect to the GDPR. Alhazmi and Arachchilage \([63]\) investigate the obstacles that developers encounter when implementing the privacy provisions in GDPR, including lack of familiarity with legal text. In a follow-up work, the authors \([64]\) propose a game design framework to help software developers better understand and implement privacy-related requirements in GDPR. Shastr et al. \([65]\) introduce anti-patterns, referring to practices of cloud-scale systems that serve their originally intended purpose well but hinder their compliance with GDPR. For example, such systems are developed to store personal data without having a clear timeline for deleting and sharing them with other applications. While this feature is naturally beneficial in terms of revenues from the system, it violates the storage and purpose limitations of GDPR.

We observe from the above-mentioned work that implementing the GDPR requirements in practice is challenging for developers. Even developers that are familiar with GDPR still struggle with identifying the complete set of requirements that is most relevant to their work. A compliant DPA contains a complete set of legally binding requirements that must be adhered to by both the controller and processor. Recall our example in Section 1 which describes an accounting office (the processor) that provides payroll administration services to Sefer University (the controller). In this case, the engineers working in the accounting office should specify explicit requirements to demonstrate that their payroll software complies with GDPR. Since the DPA contains statements specific to the context of this service, such requirements can be defined based on the DPA provided, assuming it has been properly verified and found to be GDPR-compliant. For example, the engineers should define a communication procedure, satisfying time requirements, in case of personal data breach (pursuant to S11 in Fig. 1), a module that enables rectifying personal data of Sefer University’s employees (data subjects) upon request (pursuant to S9), and an authorization module that allows only authorized users to access personal data (pursuant to S10). Our approach, presented in this paper, provides automated assistance to the developers and requirements engineers to verify the compliance of a DPA prior to relying on it for specifying compliance requirements for developing GDPR-compliant systems.

6.2 Elicitation of Compliance Requirements from Regulation

In the RE literature, representing requirements in regulations (including GDPR) relies heavily on model-driven engineering \([66]–[74]\). Other methods for representing compliance requirements have however been explored. Ayala-Rivera and Pasquale \([77]\) propose GuideMe, a systematic approach that facilitates the elicitation of requirements linking GDPR data protection obligations with privacy concerns that should be implemented in a software system. In an early work, Breaux et al. \([78]\) extract and prioritize rights and obligations from the US healthcare regulation and represent them by applying semantic parameterization, a method that enables expressing rights and obligations in restricted natural language statements. The authors derive from the regulation constraints and obligations where each phrase is attributed to an element in their representation. Following this, Breaux and Anton’s \([79]\) propose a frame-based requirements analysis method that extracts semiformal representations of requirements from regulations. More recently, formal languages \([80]–[83]\) and predefined templates \([84]\) have also been introduced in the literature to represent compliance requirements. To our knowledge, the only work to date that addresses the compliance of DPAs against GDPR is the work by Amaral et al. \([75]\). The
authors create a conceptual model that characterizes the DPA-compliance related information content according to GDPR provisions.

Compared with Amaral et al.’s work, we go one step further and present automated support for verifying the compliance of DPAs against GDPR. Another major difference (also compared to the above listed work) is that we define DPA-related compliance requirements as “shall” requirements. Our choice is motivated by the familiarity of both requirements engineers and legal experts with this format. Requirements engineers know from the common templates in RE (e.g., Rupp’s [85] and EARS [86]) that a modal verb (“shall” in this case) is an essential element indicating the importance of that requirement. Legal experts are also familiar with this format since “shall” is often used in drafting requirements in regulations [87].

The use of SRs (also referred to as semantic metadata in the RE literature) has been also discussed in the context of regulatory compliance. Semantic metadata is a prerequisite for deriving compliance requirements [52], [67], [88]. It further facilitates transforming legal text to formal specifications [89]. Several strands of work define legal concepts pertinent to requirements in regulations as semantic metadata types. For a detailed discussion on semantic metadata in the RE literature, we refer the reader to the work of Sleimi et al. [23]. In our work, we refine and adapt a set of 10 SRs from literature [22], [90]–[95].

6.3 Automated Support for Compliance Checking
Automated and semi-automated approaches using machine learning (in combination with NLP) [76], [96]–[101], as well as manual approaches through crowd-sourcing [88], [102]–[104], have been studied in the RE literature for classifying the textual content according to privacy regulations. With respect to GDPR, automated compliance has been widely studied [105]–[109]. Several approaches rely on semantic web as an enabler [110], [111]. Li et al. [112] identify a set of operationalized GDPR principles and further develop a tool to test these derived GDPR privacy requirements. The authors apply design science and their findings reveal that GDPR can be operationalized and tested through automated means. Nazarenko et al. [113] propose enriching legal text with semantic information both at a sentence-level (e.g., annotating that a sentence contains an exception) and at a phrasal-level (e.g., annotating that a phrase represents a legal entity) to facilitate semantic search and querying legal text. The authors apply their proposed approach on the French version of GDPR. Sleimi et al. [23] propose an automated rule-based method over the NLP syntax parsing to extract legal semantic metadata, e.g., an actor or a sanction, from regulations. They evaluate their approach on traffic law. Bhatia et al. [24] propose using semantic frames to detect incompleteness in privacy policies. Specifically, the authors manually analyze 15 privacy policies and conclude a set of 17 semantic roles that are expected to be present in other policies from the same domain. The authors use these SRs to further study the effect of an incomplete privacy statement (i.e., a statement that misses semantic roles) on the user’s perspective of a privacy risk.

The approaches of Sleimi et al. [23] and Bhatia et al. [24] are the closest to our work. In contrast, our work utilizes automatically-generated semantic frames for enabling automated compliance checking of DPAs at a phrasal level.

Another strand of research focuses on checking the compliance of websites and mobile applications against applicable laws. Extensive analyses of popular websites show that the variance in how organizations interpret privacy and data protection standards can lead to legal documents (e.g., privacy policies) which are not informative for individuals [114], [115]. Relevant to this, the lack of templates for drafting such legal documents can cause discrepancies between the data processing activities and what is actually disclosed to individuals [116], e.g., about the purpose of data processing (an essential requirement in GDPR) [117]. Several approaches are proposed to check the compliance of websites by classifying the content of their disclosed privacy policies according to the provisions of GDPR. Our work addresses three main challenges introduced in the above-listed work. First, we derive the compliance requirements from GDPR in collaboration with legal experts, which alleviates the risk of misinterpreting GDPR. We further make these requirements publicly available. Second, our work addresses the compliance of DPAs, another essential source for requirements for data processing activities in software systems. Our proposed automation is applicable to any DPA and does not assume conformance with any predefined template.

7 Threats to Validity
Below, we discuss threats to the validity of our empirical results and what we did to mitigate these threats.

Internal Validity. The main concern with respect to internal validity is bias. To mitigate this threat, we curated our evaluation dataset exclusively through third-parties (non-authors). The annotators were never exposed to the implementation details of our approach. Another potential threat is subjectivity in our interpretation of the GDPR text when extracting the compliance requirements related to DPA. To mitigate this threat, requirements extraction was done in close collaboration with three legal experts, who have expertise in European and international laws. Our extracted requirements are further made publicly available and thus open to scrutiny. Another threat to internal validity is our reliance on most-frequent word sense (MFS) disambiguation for identifying the semantically related words to enrich the texts of DPAs. Using more advanced disambiguation methods might improve the results of our approach. However, the wide use of MFS in the NLP literature and its reported performance provide reasonable confidence about the results of our approach. That said, further experimentation can help mitigate this threat.

External Validity. We evaluated our approach on 30 real DPAs from different sectors. Our approach performed well when identifying violations related to the GDPR requirements. This provides us with reasonable confidence that our solution is generalizable. That said, experimenting with additional DPAs can help further mitigate support external validity.

8 Extending our Methodology beyond GDPR
Our suggested procedure for assessing the compliance of DPAs can be applied on regulations beyond GDPR or document types other than DPAs. To illustrate this point, assume that a regulation is denoted as $G$, and a legal document type is denoted as $T$. In our context, GDPR is an instance of $G$, and DPAs are instances of $T$. To reuse the same methodology for other instances of $G$ and $T$, based on our experience, we anticipate the following steps and effort needed for each step.

1) Requirements elicitation from $G$ with respect to $T$, consuming about 30% of the effort. This step depends mainly on the availability of legal experts and the size and complexity of the provisions in $G$.

2) Define semantic frames (SFs) over the compliance requirements derived from $G$, consuming about 10% of the effort. Prior to defining the SFs, one has to verify whether the same semantic roles (SRs) defined in our work can be re-used as-is in the new context. If not, SRs must be refined accordingly. Defining the SFs should preferably be performed in collaboration with legal experts. Therefore, this step depends as well on their availability, though their degree of involvement would typically be much lower than in the previous step.

3) Curate an annotated dataset for $T$, consuming about 10% of the effort. The first thing to consider is whether (unlabeled) data for $T$ are available. Curating a labeled dataset of examples from $T$ is a prerequisite for developing any automation. The annotation task involves manually checking the compliance of the textual content of $T$ against the requirements derived from $G$. This step largely depends on data availability as well as the availability of annotators who have the right expertise. One should anticipate for providing training material about compliance with respect to $G$.

4) Develop an automation strategy for checking the compliance of $T$ against $G$, consuming about 50% of the effort. The main things to adjust in our approach include the extraction rules of SRs, and the similarity thresholds applied for matching the predicates and arguments between the text of $T$ and the requirements of $G$.

9 Conclusion

In this paper, we propose an automated approach for checking the compliance of data processing agreements (DPAs) against the general data protection regulation (GDPR). Such DPAs have a significant impact on the requirements of systems processing personal data. In close collaboration with legal experts, we extract and document DPA-related requirements as “shall” requirements. Such documentation provides a shared understanding between requirements engineers and legal experts. In our approach, we first manually create a representation of the GDPR requirements based on semantic frames (SFs). Our approach then automatically generates SF-based representations for the textual content of a DPA, and subsequently compares this representation against GDPR to check whether the DPA is compliant. Our evaluation is based on 30 real DPAs, curated by trained third-party annotators with a strong background in law. Over this evaluation dataset, DERECHA achieves an average accuracy of 84.6%, a precision of 89.1% and a recall of 82.4%. We also show how higher accuracy can be achieved by focused reviews on a small percentage of DPA statements. We compare DERECHA against a baseline that relies on an off-the-shelf NLP pipeline. DERECHA outperforms this baseline with an average gain of $\approx20$ percentage points in accuracy. In the future, we plan to curate more DPAs and further investigate the applicability of machine learning methods for enabling automated compliance checking. Another direction that we would like to explore is to run a large-scale experiment to verify the compliance of the DPAs available from major service providers against GDPR.

Acknowledgments

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 and CRC programs. The authors would like to specially thank Katrien Baetens, Sylvie Forastier, Peter Goes and Catherine Freichel from Linklaters LLP for enriching this work through their expertise.

References

[1] N. Zhan, S. Sarkadi, N. C. Pacheco, and J. Such, “A model for governing information sharing in smart assistants,” in AAAI/ACM Conference on AI, Ethics, and Society. AAAI Press, 2022.

[2] ´A. Feal, J. Gamba, N. Vallina-Rodriguez, P. Wijesekera, J. Reardon, S. Egelman, J. Tapiador et al., “Don’t accept candies from strangers: An analysis of third-party SDKs,” in Computers, Privacy and Data Protection Conference, 2020.

[3] European Union, “The GDPR: New opportunities, new obligations,” Justice and Consumers, 2018.

[4] N. Pantlin, C. Wiseman, and M. Everett, “Supply chain arrangements: The abc to gdpr compliance—a spotlight on emerging market practice in supplier contracts in light of the gdpr,” Computer law & Security review, vol. 34, no. 4, pp. 881–885, 2018.

[5] D. Jurafsky and J. Martin, Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, 2nd ed. Prentice Hall, 2009.

[6] European Union, “General data protection regulation,” Official Journal of the European Union, 2018.

[7] C. F. Baker, C. J. Fillmore, and J. B. Lowe, “The Berkeley frame net project,” in 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1, 1998, pp. 86–90.

[8] P. R. Kingsbury and M. Palmer, “From treebank to propbank.” in LREC, 2002, pp. 1989–1993.

[9] D. Gildea and D. Jurafsky, “Automatic labeling of semantic roles,” Computational linguistics, vol. 28, no. 3, pp. 245–288, 2002.

[10] Z. Li, H. Zhao, S. He, and J. Cai, “Syntax role for neural semantic role labeling,” Computational Linguistics, vol. 47, no. 3, pp. 529–574, 2021.

[11] L. Márquez, X. Carreras, K. C. Litojowski, and S. Stevenson, “Semantic role labeling: an introduction to the special issue,” Computational Linguistics, vol. 34, no. 2, pp. 145–159, 2008.

[12] C. Shi, S. Liu, S. Ren, S. Feng, M. Li, M. Zhou, X. Sun, and H. Wang, “Knowledge-based semantic embedding for machine translation,” in Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2016, pp. 2245–2254.

[13] J. Berant, A. Chou, R. Frostig, and P. Liang, “Semantic parsing on freebase from question-answer pairs,” in Proceedings of the 2013 conference on empirical methods in natural language processing, 2013, pp. 1533–1544.

[14] D. Shen and M. Lapata, “Using semantic roles to improve question answering,” in Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning, 2007, pp. 12–21.
[103] S. Wilson, F. Schaub, R. Ramanath, N. M. Sadeh, F. Liu, N. A. Smith, and F. Liu, “Crowdsourcing annotations for websites’ privacy policies: Can it really work?” in *Proceedings of the 25th International Conference on World Wide Web*, 2016.

[104] T. Linden, R. Khandelwal, H. Harkous, and K. Fawaz, “The privacy policy landscape after the gdpr,” *Proceedings on Privacy Enhancing Technologies*, vol. 2020, no. 1, pp. 47–64, 2020.

[105] E. Park, S. Han, H. Bae, R. Kim, S. Lee, D. Lim, and H. Lim, “Development of automatic evaluation tool for mobile accessibility for android application,” in *2019 International Conference on Systems of Collaboration Big Data, Internet of Things & Security (SysCoBioTS)*, 2019, pp. 1–6.

[106] J. Benjumea, E. Dorronzoro, J. Ropero, O. Rivera-Romero, and A. Carrasco, “Privacy in mobile health applications for breast cancer patients,” in *2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)*, 2019, pp. 634–639.

[107] L. Verderame, D. Caputo, A. Romdhana, and A. Merlo, “On the (un)reliability of privacy policies in android apps,” in *2020 International Joint Conference on Neural Networks (IJCNN)*, 2020, pp. 1–9.

[108] D. S. Guaman, J. M. Del Alamo, and J. C. Caiza, “Gdpr compliance assessment for cross-border personal data transfers in android apps,” *IEEE Access*, vol. 9, pp. 15,961–15,982, 2021.

[109] M. Cozar, D. Rodriguez, J. M. Del Alamo, and D. Guaman, “Reliability of ip geolocation services for assessing the compliance of international data transfers,” in *2022 IEEE European Symposium on Security and Privacy Workshops on Privacy in the Electronic Society (EuropesPW)*, 2022, pp. 181–185.

[110] M. Palmirani and G. Governatori, “Modelling legal knowledge for GDPR compliance checking,” in *Legal Knowledge and Information Systems: The Thirty-first Annual Conference, Groningen, The Netherlands, 12-14 December 2018*, ser. Frontiers in Artificial Intelligence and Applications, M. Palmirani, Ed., vol. 313. IOS Press, 2018, pp. 101–110.

[111] P. A. Bonatti, S. Kirrane, I. M. Petrova, and L. Sauro, “Machine understandable policies and gdpr compliance checking,” *Künstliche Intelligenz*, vol. 34, no. 3, pp. 303–315, 2020.

[112] Z. S. Li, C. Werner, N. Ernst, and D. Damian, “Gdpr compliance in the context of continuous integration,” *arXiv preprint arXiv:2002.06830*, 2020.

[113] A. Nazarenko, F. Lévy, and A. Wyner, “A pragmatic approach to semantic annotation for search of legal texts - an experiment on gdpr,” in *Legal Knowledge and Information Systems: The Thirty-fourth Annual Conference, Vilnius, Lithuania, 8-10 December 2021*, ser. Frontiers in Artificial Intelligence and Applications, S. Erich, Ed., vol. 346. IOS Press, 2021, pp. 23–32.

[114] R. Chen, F. Fang, T. Norton, A. M. McDonald, and N. Sadeh, “Fighting the fog: Evaluating the clarity of privacy disclosures in the age of ccpa,” in *Proceedings of the 20th Workshop on Privacy in the Electronic Society*, 2021, pp. 73–102.

[115] M. Degeling, C. Utz, C. Lentzsch, H. Hosseini, F. Schaub, and T. Holz, “We value your privacy... now take some cookies: Measuring the gdpr’s impact on web privacy,” *Informatik Spektrum*, vol. 42, 2019.

[116] M. S. Rahman, P. Naghavi, B. Kojsner, S. Afroz, B. Williams, S. Rampazzi, and V. Bindschaedler, “Permpress: Machine learning-based pipeline to evaluate permissions in app privacy policies,” *IEEE Access*, vol. 10, 2022.

[117] A.-J. Aberkane, S. V. Broucke, and G. Poels, “Investigating organizational factors associated with gdpr noncompliance using privacy policies: A machine learning approach,” in *2022 IEEE 4th International Conference on Trust, Privacy and Security in Intelligent Systems, and Applications (TPS-ISA)*. IEEE, 2022, pp. 107–113.
Orlando Amaral Cejas obtained his B.Sc. in Automation Engineering (2012) and his M.Sc. in Digital Systems (2017) at the Technological University of Havana (Cuba). During that time, he participated in the improvement of the Cuban intelligent transport system. Currently, he is working as a doctoral researcher at SnT centre for Security, Reliability and Trust at the University of Luxembourg. He is doing his PhD as part of an industrial project on compliance checking of legal text according to the GDPR. His main research interests include: regulatory compliance, model-driven engineering, applied natural language processing, text mining and machine learning.

Muhammad Ilyas Azeem is working as a research associate at the SnT centre for Security, Reliability, and Trust, University of Luxembourg. He obtained his PhD in Software Engineering at the Institute of Software, University of Chinese Academy of Sciences, Beijing, China in 2020. He is currently working on an industrial project on compliance checking of legal text according to the GDPR. His main research interests include requirement engineering, mining software repositories, natural language processing, machine learning, language models, and regulatory compliance.

Sallam Abualhaija is a research scientist at the Interdisciplinary Centre for Security, Reliability, and Trust (SnT), University of Luxembourg. She received her PhD in Computer Science (2016) at Hamburg University of Technology (Germany). Her main research interests combine AI with software engineering, with an emphasis on: requirements engineering, regulatory compliance, natural language processing, and machine learning. Abualhaija believes in fostering stronger ties between academia and industry and has conducted the majority of her research in close collaboration with industry partners in the past years.

Lionel C. Briand is professor of software engineering and has shared appointments between (1) School of Electrical Engineering and Computer Science, University of Ottawa, Canada and (2) The SnT centre for Security, Reliability, and Trust, University of Luxembourg. He is the head of the SVV department at the SnT Centre and a Canada Research Chair in Intelligent Software Dependability and Compliance (Tier 1). He has conducted applied research in collaboration with industry for more than 25 years, including projects in the automotive, aerospace, manufacturing, financial, and energy domains. He is a fellow of the IEEE and ACM. He was also granted the IEEE Computer Society Harlan Mills award (2012), the IEEE Reliability Society Engineer-of-the-year award (2013), and the ACM SIGSOFT Outstanding Research Award (2022) for his work on software testing and verification. More details can be found on: http://www.lbriand.info.
### APPENDIX A

**DPA Compliance Requirements in GDPR**

| ID (Cat<sup>1</sup>) | Requirement (Reference<sup>2</sup>) |
|----------------------|-------------------------------------|
| R27 (MD7)            | The organizational and technical measures to ensure a level of security can include: (a) pseudonymization and encryption of personal data, (b) ensure confidentiality, integrity, availability and resilience of processing systems and services, (c) restore the availability and access to personal data in a timely manner in the event of a physical or technical incident, and (d) regularly testing, assessing and evaluating the effectiveness of technical and organizational measures for ensuring the security of the processing. (*Art. 32(1)*) |
| R28 (MD8)            | The notification of personal data breach shall at least include (a) the nature of personal data breach; (b) the name and contact details of the data protection officer; (c) the consequences of the breach; (d) the measures taken or proposed to mitigate its effects. (*Art. 33(3)*) |
| R29 (MD9)            | The DPIA shall at least include (a) a systematic description of the envisaged processing operations and the purposes of the processing, (b) an assessment of the necessity and proportionality of the processing operations in relation to the purposes, (c) an assessment of the risks to the rights and freedoms of data subjects, and (d) the measures envisaged to address the risks. (*Art. 35(7)*) |
| R30 (PO20)           | The processor shall not transfer personal data to a third country or international organization without a prior specific or general authorization of the controller. (*Art. 28(3)(a)*) |
| R31 (PO21)           | The processor can demonstrate guarantees to Article 28 (1–4) through adherence to an approved codes of conduct or an approved certification mechanism. (*Art. 28(5)*) |
| R32 (PO22)           | The processor shall implement appropriate technical and organizational measures to ensure a level of security appropriate to the risk of varying likelihood and severity for the rights and freedoms of natural persons. (*Art. 32(1)*) |
| R33 (PO23)           | The processor shall ensure that any natural person acting under its authority who has access to personal data only process them on instructions from the controller. (*Art. 32(4)*) |
| R34 (PO24)           | The processor shall notify the controller without undue delay after becoming aware of a personal data breach. (*Art. 33(2)*) |
| R35 (PO25)           | A processor shall be liable for the damage caused by processing only where it has not complied with GDPR obligations specifically directed to processors or where it has acted outside or contrary to lawful instructions of the controller. (*Art. 82(2)*) |
| R36 (CO1)            | The controller shall inform the supervisory authority no later than 72 hours after having become aware of a personal data breach. (*Art. 33(1)*) |
| R37 (CO2)            | The controller shall document personal data breaches. (*Art. 33(5)*) |
| R38 (CO3)            | In case of high risks, the controller shall communicate the data breach to the data subject without undue delay. (*Art. 34(1)*) |
| R39 (CO4)            | The controller shall carry out a DPIA. (*Art. 35(1)*) |
| R40 (CO5)            | The controller shall seek advice of the DPO when carrying a DPIA. (*Art. 35(2)*) |
| R41 (CO6)            | The controller shall seek the views of data subjects or their representatives on the intended processing. (*Art. 35(9)*) |
| R42 (CO7)            | The controller shall carry out a review to assess if processing is performed in accordance with the data protection impact assessment at least when there is a change of the risk represented by processing operations. (*Art. 35(11)*) |
| R43 (CO8)            | A controller shall be liable for the damage caused by any processing infringing the GDPR. (*Art. 82(2)*) |
| R44 (CR2)            | The controller shall have the right to suspend the processing in certain cases. (*Linklaters LLP*) |
| R45 (CR3)            | The controller shall have the right to terminate the DPA in certain cases. (*Linklaters LLP*) |

<sup>1</sup> Cat is the category of the requirement, namely metadata (MD), processor’s obligation (PO), controller’s right (CR), and controller’s obligation (CO).

<sup>2</sup> GDPR-related articles.

You can choose not to have a title for an appendix if you want by leaving the argument blank.
TABLE 8: Glossary Terms for DPA Compliance Requirements in GDPR.

| Concept (Art. 1)                              | Definition                                                                 |
|-----------------------------------------------|---------------------------------------------------------------------------|
| Personal Data (Art. 4(1))                    | Any information relating to an identified or identifiable natural person (data subject). |
| Processing (Art. 4(2))                       | Any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organization, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction. |
| Pseudonymization (Art. 4(5))                 | The processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organizational measures to ensure that the personal data are not attributed to an identified or identifiable natural person. |
| Data Controller (Art. 4(7))                  | A natural or legal person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data; where the purposes and means of such processing are determined by Union or Member State law, the controller or the specific criteria for its nomination may be provided for by Union or Member State law. |
| Data Processor (Art. 4(8))                   | A natural or legal person, public authority, agency or other body which processes personal data on behalf of the controller. |
| Personal Data Breach (Art. 4(12))            | A breach of security leading to the accidental or unlawful destruction, loss, alteration, unauthorized disclosure of, or access to, personal data transmitted, stored or otherwise processed. |
| Supervisory Authority (Art. 4(22))           | A supervisory authority which is concerned by the processing of personal data because the controller or processor is established on the territory of the Member State of that supervisory authority; data subjects residing in the Member State of that supervisory authority are substantially affected or likely to be substantially affected by the processing; or a complaint has been lodged with that supervisory authority. |
| Objection (Art. 4(24))                       | An objection to a draft decision as to whether there is an infringement of this Regulation, or whether envisaged action in relation to the controller or processor complies with this Regulation, which clearly demonstrates the significance of the risks posed by the draft decision as regards the fundamental rights and freedoms of data subjects and, where applicable, the free flow of personal data within the Union. |
| International Organization (Art. 4(26))       | An organization and its subordinate bodies governed by public international law, or any other body which is set up by, or on the basis of, an agreement between two or more countries. |
| Sub-processor (Art. 28(4))                   | A natural person or organization engaged by a Processor for carrying out specific processing activities on behalf of the Controller. |
| Security of Processing (Art. 32(1))          | Taking into account the state of the art, the costs of implementation and the nature, scope, context and purposes of processing as well as the risk of varying likelihood and severity for the rights and freedoms of natural persons, the controller and the processor shall implement appropriate technical and organizational measures to ensure a level of security appropriate to the risk. |
| Data Protection Impact Assessment (Art. 35(7))| An assessment of the necessity and proportionality of the processing operations in relation to the purposes, considering the risks to the rights, freedoms and legitimate interests of data subjects (abbreviated as DPIA). |
| Codes of Conduct (Art. 40(1))                | Are intended to contribute to the proper application of this Regulation, taking account of the specific features of the various processing sectors and the specific needs of micro, small and medium-sized enterprises. |
| Certification (Art. 42(1))                   | The establishment of data protection certification mechanisms and of data protection seals and marks, for the purpose of demonstrating compliance with this Regulation of processing operations by controllers and processors. The specific needs of micro, small and medium-sized enterprises shall be taken into account. |
| Personal Data Transfer (Art. 45(1))          | A transfer of personal data to a third country or an international organization may take place where the Commission has decided that the third country, a territory or one or more specified sectors within that third country, or the international organization in question ensures an adequate level of protection. If agreed by the parties, for the transfer of personal data processed under this agreement from a third country or international organization, they can rely on different mechanisms, including the prior consent of the controller. |

1 GDPR-related articles.
### Appendix B

**SF-based Representations of Mandatory and Optional Requirements**

| ID   | Representation (p), A | Conditions                                                                 |
|------|-----------------------|-----------------------------------------------------------------------------|
| PO1  | (not engage), A = \{a_i: 0 \leq i \leq 3\} such that \(a_0 = \) the processor\textsubscript{actor}, \(a_1 = \) a sub-processor\textsubscript{object}, \(a_2 = \) [without prior specific or general written approval of the controller]\textsubscript{constraint}, \(a_3 = \) [prior specific or general written approval]\textsubscript{time} |                                                                                   |
| PO2  | (inform), A = \{a_i: 0 \leq i \leq 4\} such that \(a_0 = \) the processor\textsubscript{actor}, \(a_1 = \) [the controller]\textsubscript{beneficiary}, \(a_2 = \) [any intended changes]\textsubscript{object}, \(a_3 = \) [the addition or replacement of sub-processors]\textsubscript{situation}, \(a_4 = \) [in case of written authorization]\textsubscript{constraint} |                                                                                   |
| PO3  | (process), A = \{a_i: 0 \leq i \leq 2\} such that \(a_0 = \) the processor\textsubscript{actor}, \(a_1 = \) [personal data]\textsubscript{object}, \(a_2 = \) [on documented instructions from the controller]\textsubscript{constraint} |                                                                                   |
| PO4  | (inform), A = \{a_i: 0 \leq i \leq 5\} such that \(a_0 = \) the processor\textsubscript{actor}, \(a_1 = \) [that legal requirement]\textsubscript{object}, \(a_2 = \) [the controller]\textsubscript{beneficiary}, \(a_3 = \) [If the processor requires by Union or Member State law to process personal data without instructions and law does not prohibit informing the controller on grounds of public interest]\textsubscript{condition}, \(a_4 = \) [before processing]\textsubscript{time}, \(a_5 = \) [Union or Member State law]\textsubscript{reference} |                                                                                   |
| PO5  | (ensure), A = \{a_i: 0 \leq i \leq 3\} such that \(a_0 = \) the processor\textsubscript{actor}, \(a_1 = \) [persons]\textsubscript{object}, \(a_2 = \) [have committed themselves to confidentiality]\textsubscript{situation}, \(a_3 = \) [are under an appropriate statutory obligation of confidentiality]\textsubscript{constraint} |                                                                                   |
| PO6  | (take), A = \{a_i: 0 \leq i \leq 3\} such that \(a_0 = \) the processor\textsubscript{actor}, \(a_1 = \) [all measures]\textsubscript{object}, \(a_2 = \) [Article 32]\textsubscript{reference}, \(a_3 = \) [to ensure the security of processing]\textsubscript{reason} |                                                                                   |
| PO7  | (assist), A = \{a_i: 0 \leq i \leq 4\} such that \(a_0 = \) the processor\textsubscript{actor}, \(a_1 = \) [the controller]\textsubscript{beneficiary}, \(a_2 = \) [in fulfilling its obligation]\textsubscript{object}, \(a_3 = \) [to respond to requests for exercising the data subject’s rights]\textsubscript{reason}, \(a_4 = \) [requests]\textsubscript{situation} |                                                                                   |
| PO8  | (assist), A = \{a_i: 0 \leq i \leq 2\} such that \(a_0 = \) the processor\textsubscript{actor}, \(a_1 = \) [the controller]\textsubscript{beneficiary}, \(a_2 = \) [in ensuring the security of processing]\textsubscript{object} |                                                                                   |
| PO9  | (assist), A = \{a_i: 0 \leq i \leq 5\} such that \(a_0 = \) the processor\textsubscript{actor}, \(a_1 = \) [the controller]\textsubscript{beneficiary}, \(a_2 = \) [in consulting the supervisory authorities]\textsubscript{object}, \(a_3 = \) [prior to processing]\textsubscript{time}, \(a_4 = \) [to mitigate the risk]\textsubscript{reason}, \(a_5 = \) [where the processing would result in a high risk in the absence of measures taken by the controller]\textsubscript{constraint} |                                                                                   |
| PO10 | (assist), A = \{a_i: 0 \leq i \leq 3\} such that \(a_0 = \) the processor\textsubscript{actor}, \(a_1 = \) [the controller]\textsubscript{beneficiary}, \(a_2 = \) [in notifying to the supervisory authority]\textsubscript{object}, \(a_3 = \) [a personal data breach]\textsubscript{situation} |                                                                                   |
| PO11 | (assist), A = \{a_i: 0 \leq i \leq 3\} such that \(a_0 = \) the processor\textsubscript{actor}, \(a_1 = \) [the controller]\textsubscript{beneficiary}, \(a_2 = \) [in communicating to the data subject]\textsubscript{object}, \(a_3 = \) [a personal data breach]\textsubscript{situation} |                                                                                   |
| PO12 | (assist), A = \{a_i: 0 \leq i \leq 2\} such that \(a_0 = \) the processor\textsubscript{actor}, \(a_1 = \) [the controller]\textsubscript{beneficiary}, \(a_2 = \) [in ensuring compliance with the obligations pursuant to data protection impact assessment (DPIA)]\textsubscript{object} |                                                                                   |
| PO13 | (return or delete), A = \{a_i: 0 \leq i \leq 4\} such that \(a_0 = \) the processor\textsubscript{actor}, \(a_1 = \) [all personal data]\textsubscript{object}, \(a_2 = \) [the controller]\textsubscript{beneficiary}, \(a_3 = \) [after end]\textsubscript{time}, \(a_4 = \) [the end of the provision of services related to processing]\textsubscript{situation} |                                                                                   |
| PO14 | (inform), A = \{a_i: 0 \leq i \leq 3\} such that \(a_0 = \) the processor\textsubscript{actor}, \(a_1 = \) [the controller]\textsubscript{beneficiary}, \(a_2 = \) [if an instruction infringes the GDPR or other data protection provisions]\textsubscript{condition}, \(a_3 = \) [GDPR or other data protection provisions]\textsubscript{reference} |                                                                                   |
| PO15 | (make available), A = \{a_i: 0 \leq i \leq 2\} such that \(a_0 = \) the processor\textsubscript{actor}, \(a_1 = \) [the controller]\textsubscript{beneficiary}, \(a_2 = \) [information necessary to demonstrate compliance with the obligations]\textsubscript{object}, \(a_3 = \) [Article 28 in GDPR]\textsubscript{reference} |                                                                                   |
| PO16 | (allow for and contribute to), A = \{a_i: 0 \leq i \leq 2\} such that \(a_0 = \) the processor\textsubscript{actor}, \(a_1 = \) [audits including inspections]\textsubscript{object}, \(a_2 = \) [conducted by the controller or another auditor mandated by the controller]\textsubscript{situation} |                                                                                   |
| PO17 | (impose), A = \{a_i: 0 \leq i \leq 4\} such that \(a_0 = \) the processor\textsubscript{actor}, \(a_1 = \) [the engaged sub-processors]\textsubscript{beneficiary}, \(a_2 = \) [the same obligations]\textsubscript{object}, \(a_3 = \) [by way of contract or other legal act under]\textsubscript{constraint}, \(a_4 = \) [Union or Member State law]\textsubscript{reference} |                                                                                   |
| PO18 | (remain fully liable), A = \{a_i: 0 \leq i \leq 2\} such that \(a_0 = \) the processor\textsubscript{actor}, \(a_1 = \) [the controller]\textsubscript{beneficiary}, \(a_2 = \) [for the performance of sub-processor’s obligations]\textsubscript{object} |                                                                                   |
| PO19 | (take into account), A = \{a_i: 0 \leq i \leq 2\} such that \(a_0 = \) the processor\textsubscript{actor}, \(a_1 = \) [the risk of]\textsubscript{object}, \(a_2 = \) [accidental or unlawful destruction, loss, alteration, unauthorized disclosure of or access to the personal data transmitted, stored or processed]\textsubscript{situation} |                                                                                   |
| CR1  | (have the right), A = \{a_i: 0 \leq i \leq 5\} such that \(a_0 = \) [the controller]\textsubscript{actor}, \(a_1 = \) [to object to changes]\textsubscript{object}, \(a_2 = \) [in case of written authorization]\textsubscript{condition}, \(a_3 = \) [by the processor]\textsubscript{constraint}, \(a_4 = \) [after having been informed of such intended changes]\textsubscript{time}, \(a_5 = \) [addition or replacement of sub-processors]\textsubscript{situation} |                                                                                   |
| ID  | Representation (p. A)                                                                 |
|-----|---------------------------------------------------------------------------------------|
| PO20 | (not transfer), A = \{a_i : 0 \leq i \leq 4\} such that a_0 = [the processor]_{\text{actor}}, a_1 = [\text{a third country or international organization}]_{\text{beneficiary}}, a_2 = [\text{personal data}]_{\text{object}}, a_3 = [\text{without a prior specific or general authorization of the controller}]_{\text{constraint}}, a_4 = [\text{prior specific or general authorization}]_{\text{time}} |
| PO21 | (demonstrate), A = \{a_i : 0 \leq i \leq 3\} such that a_0 = [\text{the processor}]_{\text{actor}}, a_1 = [\text{guarantees}]_{\text{subject}}, a_2 = [\text{adherence to an approved code of conduct or an approved certification mechanism}]_{\text{situation}}, a_3 = [\text{Article 28 (1-4)}]_{\text{reference}} |
| PO22 | (implement), A = \{a_i : 0 \leq i \leq 3\} such that a_0 = [\text{the processor}]_{\text{actor}}, a_1 = [\text{appropriate technical and organizational measures}]_{\text{object}}, a_2 = [\text{to ensure a level of security appropriate to the risk}]_{\text{reason}}, a_3 = [\text{varying likelihood and severity for the rights and freedoms of natural persons}]_{\text{situation}} |
| PO23 | (ensure), A = \{a_i : 0 \leq i \leq 3\} such that a_0 = [\text{the processor}]_{\text{actor}}, a_1 = [\text{any person}]_{\text{beneficiary}}, a_2 = [\text{under its authority who has access to personal data acts only on instructions from the controller}]_{\text{constraint}}, a_3 = [\text{access to personal data}]_{\text{situation}} |
| PO24 | (notify), A = \{a_i : 0 \leq i \leq 3\} such that a_0 = [\text{the processor}]_{\text{actor}}, a_1 = [\text{the controller}]_{\text{beneficiary}}, a_2 = [\text{without undue delay}]_{\text{constraint}}, a_3 = [\text{a data breach}]_{\text{situation}} |
| PO25 | (be liable), A = \{a_i : 0 \leq i \leq 3\} such that a_0 = [\text{a processor}]_{\text{actor}}, a_1 = [\text{for the damage caused}]_{\text{object}}, a_2 = [\text{where acting outside or contrary to lawful instructions of the controller or not complying with obligations of the GDPR specifically directed to processors}]_{\text{condition}}, a_3 = [\text{by processing}]_{\text{constraint}} |
| CO1  | (inform), A = \{a_i : 0 \leq i \leq 3\} such that a_0 = [\text{the controller}]_{\text{actor}}, a_1 = [\text{the supervisory authority}]_{\text{beneficiary}}, a_2 = [\text{no later than 72 hours after having become aware}]_{\text{constraint}}, a_3 = [\text{a personal data breach}]_{\text{situation}} |
| CO2  | (document), A = \{a_i : 0 \leq i \leq 1\} such that a_0 = [\text{the controller}]_{\text{actor}}, a_1 = [\text{personal data breaches}]_{\text{object}} |
| CO3  | (inform), A = \{a_i : 0 \leq i \leq 4\} such that a_0 = [\text{the controller}]_{\text{actor}}, a_1 = [\text{the data subject}]_{\text{beneficiary}}, a_2 = [\text{in case of high risks}]_{\text{condition}}, a_3 = [\text{without undue delay}]_{\text{constraint}}, a_4 = [\text{a data breach}]_{\text{situation}} |
| CO4  | (carry out), A = \{a_i : 0 \leq i \leq 1\} such that a_0 = [\text{the controller}]_{\text{actor}}, a_1 = [\text{DPIA}]_{\text{object}} |
| CO5  | (seek), A = \{a_i : 0 \leq i \leq 2\} such that a_0 = [\text{the controller}]_{\text{actor}}, a_1 = [\text{advice of the DPO}]_{\text{object}}, a_2 = [\text{when carrying out DPIA}]_{\text{condition}} |
| CO6  | (seek), A = \{a_i : 0 \leq i \leq 1\} such that a_0 = [\text{the controller}]_{\text{actor}}, a_1 = [\text{the views of data subjects or their representatives on the intended processing}]_{\text{object}} |
| CO7  | (carry out), A = \{a_i : 0 \leq i \leq 5\} such that a_0 = [\text{the controller}]_{\text{actor}}, a_1 = [\text{a review}]_{\text{object}}, a_2 = [\text{at least when there exists}]_{\text{condition}}, a_3 = [\text{represented by processing operations}]_{\text{constraint}}, a_4 = [\text{to assess if processing is performed in accordance with the DPIA}]_{\text{reason}}, a_5 = [\text{a change of the risk}]_{\text{situation}} |
| CO8  | (be liable), A = \{a_i : 0 \leq i \leq 3\} such that a_0 = [\text{a controller}]_{\text{actor}}, a_1 = [\text{for the damage}]_{\text{object}}, a_2 = [\text{caused by any processing infringing the GDPR}]_{\text{constraint}}, a_3 = [\text{GDPR}]_{\text{reference}} |
| CR2  | (have the right), A = \{a_i : 0 \leq i \leq 2\} such that a_0 = [\text{the controller}]_{\text{actor}}, a_1 = [\text{to suspend the processing}]_{\text{object}}, a_2 = [\text{in certain cases}]_{\text{condition}}, a_3 = [\text{the DPA}]_{\text{reference}} |
| CR3  | (have the right), A = \{a_i : 0 \leq i \leq 3\} such that a_0 = [\text{the controller}]_{\text{actor}}, a_1 = [\text{to terminate}]_{\text{object}}, a_2 = [\text{in certain cases}]_{\text{condition}}, a_3 = [\text{the DPA}]_{\text{reference}} |