IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 5, 2021, accepted April 23, 2021, date of publication May 3, 2021, date of current version May 10, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3076921

Exploring Automated GDPR-Compliance in Requirements Engineering: A Systematic Mapping Study

ABDEL-JAOUAD ABERKANE, GEERT POELS, AND SEPPE VANDEN BROUCKE
Business Informatics Research Group, Faculty of Economics and Business Administration, Ghent University, 9000 Ghent, Belgium
Corresponding author: Abdel-Jaouad Aberkane (abdeljaouad.aberkane@ugent.be)

ABSTRACT The General Data Protection Regulation (GDPR), adopted in 2018, profoundly impacts information processing organizations as they must comply with this regulation. In this research, we consider GDPR-compliance as a high-level goal in software development that should be addressed at the outset of software development, meaning during requirements engineering (RE). In this work, we hypothesize that natural language processing (NLP) can offer a viable means to automate this process. We conducted a systematic mapping study to explore the existing literature on the intersection of GDPR, NLP, and RE. As a result, we identified 448 relevant studies, of which the majority (420) were related to NLP and RE. Research on the intersection of GDPR and NLP yielded nine studies, while 20 studies were related to GDPR and RE. Even though only one study was identified on the convergence of GDPR, NLP, and RE, the mapping results indicate opportunities for bridging the gap between these fields. In particular, we identified possibilities for introducing NLP techniques to automate manual RE tasks in the crossing of GDPR and RE, in addition to possibilities of using NLP-based machine learning techniques to achieve GDPR-compliance in RE.

INDEX TERMS General data protection regulation, systematic mapping study, requirements engineering, natural language processing.

I. INTRODUCTION

As of 25 May 2018, the General Data Protection Regulation (GDPR) came into effect to protect the processing of personal data and thus endeavoring to assure the rights of data subjects [1]. Initiated by the European Union, this development has given individuals located within the European Union more control over their data and forces organizations that fall under its jurisdiction to adhere to the GDPR, while insubordination may lead to financial penalties and possibly to loss of reputation. It is therefore imperative for organizations to consider GDPR-compliance at the outset of developing information systems. For this reason, the GDPR urges organizations to meet, in particular, the principles of data protection by design and data protection by default.

Recital 78 of the GDPR states that ''to be able to demonstrate compliance with this Regulation, the controller should adopt internal policies and implement measures which meet in particular the principles of data protection by design and data protection by default'' [1]. By addressing data protection considerations—as decreed by the GDPR—in the development process of software systems, organizations ensure awareness of the regulations among involved professionals and avoid haphazardness in designing the software system. This process of eliciting high-level goals early in the development process is crucial for solidifying these goals in the system, which brings us to the field of Requirements Engineering (RE) [2]. In this research, we approach GDPR-compliance as a high-level goal in software development, and for this reason, we consider RE as the quintessential paradigm to address the GDPR and realize data protection by design and data protection by default.

GDPR-compliance demands significant efforts in the form of substantial financial and human resources, along with training of employees [3]. To abet organizations in achieving compliance from an RE lens, automation in the form of natural language processing (NLP) may be the answer. Considering the availability of NLP tools in RE, NLP can empower RE professionals in their tasks [4]—and the organization as a whole—by adopting a more proactive stance

The associate editor coordinating the review of this manuscript and approving it for publication was Imran Sarwar Bajwa.

towards GDPR-compliance. This is especially the case when we consider the significance and increase of natural language use in requirements representation [5].

To our knowledge, no systematic mapping study combining the fields of NLP and RE while centering on GDPR-compliance has already been published. With this research, we aim to fill this gap and aid organizations in meeting the GDPR by collecting approaches and solutions in RE, NLP, or both, that can be used for GDPR-compliance. To gather this information, we pose the following three research questions and aim to answer them through a systematic mapping study: 1) What NLP approaches are useful for RE and for which activities?; 2) Which NLP approaches are available for achieving GDPR-compliance in organizations?, and; 3) Which state-of-the-art RE solutions support the GDPR-guidelines? The answers to these questions will provide us with insight into the state-of-the-art approaches that can be used for achieving automated GDPR-compliance during the RE process of software development.

The rest of this paper is structured as follows. Section 2 provides background information, related work and identifies the research gap that we address. Section 3 describes the adopted approach for our systematic mapping study and outlines the literature mapping protocol. Section 4 describes the results of the mapping study. Section 5 describes the principal findings and threats to the validity of our approach. Finally, Section 6 concludes our research and provides pointers to future work.

II. BACKGROUND

In this research, we define automated GDPR-compliance as utilizing NLP approaches in RE activities to comply with the GDPR. To achieve GDPR-compliance, companies need to invest in, inter alia, workforce, resources, and technology platforms—however, most organizations are not yet adequately prepared [6]. For this reason, we conduct a systematic mapping study to outline available RE and NLP approaches that can be used for automated GDPR-compliance, consequently researching the crossroads of GDPR, NLP, and RE. This section will discuss the relevance of these research fields, the implications of the GDPR, and finally, a brief overview of related work is given.

A. GENERAL DATA PROTECTION REGULATION

The GDPR is a European regulation that aims to protect personal data and becomes relevant for companies as soon as any data processing takes place [1]. The GDPR applies to the European Economic Area (EEA) [7] and to every organization outside the EEA that processes data of European data subjects and has, therefore, a significant impact on the digital market of the world [8]. The GDPR defines seven principles related to personal data processing: lawfulness, fairness and transparency, purpose limitation, data minimization, accuracy, storage limitation, integrity and confidentiality, and accountability. These principles yield the following data subject rights: right of access, right to be informed, right to rectification, right to erasure, right to restriction of processing, right to data portability, right to object, and rights related to automated decision-making, including profiling. Article 25 of the GDPR states that the controller—the person or entity that determines the purposes and means of the processing of personal data—should implement appropriate technical and organizational measures which are designed to enforce these data-protection principles [1]. The recognition that the conditions for data processing are fundamentally being set by the soft- and hardware used for the task [9], opens the gate for RE to enter the discourse. As will be explained in the following sub-section, RE is concerned with understanding and realizing a software system's higher goal. One does not need much imagination to identify the implementation of appropriate technical measures as an RE challenge. Hence, in this research, we will look at GDPR-compliance in the light of RE.

B. REQUIREMENTS ENGINEERING

RE entails elicitation, evaluation, specification, analysis, and evolution of the objectives, functionalities, qualities, and constraints to be achieved by a software-intensive system within some organizational or physical environment [10]. These activities collaborate in order to understand the intended higher goal of a software system by identifying stakeholders and their needs, and documenting these in a form that is amenable to analysis, communication, and subsequent implementation [2]. ''Inappropriate and ill-defined software requirements'' detriment information systems projects, and thus stress the importance of RE [11]. Due to the importance and defining character of RE, we consider this paradigm potent for instilling the GDPR in software systems. In this research, we consider GDPR-compliance as one of the higher goals of a system; assuming that the corresponding organization falls under the GDPR's jurisdiction.

C. NATURAL LANGUAGE PROCESSING

According to [12], NLP is ''a theoretically motivated range of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis'' to achieve human-like language processing. From this definition follows—as argued by the author—that NLP is not usually considered a goal in and of itself, which we aim to exemplify in this research by mapping NLP approaches useful for achieving automated GDPR-compliance in RE. RE is a particularly fertile field for the use of NLP since requirements are regularly expressed in natural language [5]. This is inherently true for GDPR data protection requirements. Furthermore, NLP technologies may come in useful by assisting software professionals in dissecting the legalese nature of the GDPR. Despite the growing interest in applying NLP techniques to RE practice [4], it may be that—due to the novelty of the GDPR—not many automated NLP approaches will be available as of yet for this cause. By investigating this, we aim to shed light and outline the NLP possibilities for automated GDPR-compliance in RE.

D. RELATED WORK

A brief review of the literature—conducted on March 2020 using the IEEE Xplore Digital Library—demonstrates that no systematic literature study is available where RE approaches and NLP approaches are aggregated for automated GPDR-compliance. Our preliminary search shows that literature in these fields (i.e., GDPR, NLP, and RE) broadly falls into two categories: RE and GDPR-compliance, and NLP for RE. Research in the first category does—to our best knowledge—not comprise literature studies, whereas literature in the second category does comprise literature studies. The next paragraphs show a preliminary review of relevant literature. These studies are summarized in Table 1.

TABLE 1. Preliminary review of related work.

Study	Research Goal or Solution Proposal	Research Domain(s)
[13]	Modelling the GDPR using UML and OCL to pave the way for developing automated model-based GDPR compliance analysis solutions.	GDPR, RE
[14]	GuideMe, a 6-step approach to elicit solution requirements that ensure compliance with the legal obligations imposed by the GDPR.	GDPR, RE
[15]	A modelling framework to support the design of GDPR compliant systems.	GDPR, RE
[16]	An approach to requirements engineering with regard to developing GDPR compliant software.	GDPR, RE
[19]	Systematic literature review of utilization of NLP in the domain of Software Requirement Engineering.	NLP, RE
[18]	Systematic literature review of NLP-techniques that address ambiguity in requirements.	NLP, RE
[20]	Systematic literature review of machine learning algorithms for identification and classification of NFRs.	NLP, RE
[17]	Systematic literature review of the literature in automated requirements elicitation.	NLP, RE
[4]	Outline future research directions of NLP in RE.	NLP, RE

Focusing on RE and the GDPR, Torre *et al.* [13] present an approach to model the GDPR using UML and OCL as a first step towards developing future model-based automated methods for assessing GDPR-compliance. Furthermore, the authors report, not without importance, that no automated approach for checking GDPR-compliance had been published at the time of publishing their article. In the same research stream, Ayala-Rivera and Pasquale [14] present GuideME, a 6-step systematic approach that supports practitioners in the elicitation of solution requirements from the GDPR legal obligations. Moreover, Robol *et al.* [15] propose a method to support the design of GDPR-compliant systems based on a socio-technical approach composed of a modeling language and a reasoning framework. Furthermore, Ringmann *et al.* [16] present a proposal of generic reusable technical requirements for the software development process that satisfies the key principles of the GDPR.

On the crossroads of NLP and RE, Meth *et al.* [17] report a systematic review of literature in automated requirements elicitation, focusing on the first activity of requirements engineering. In a different literature review, Bano [18] maps the application of NLP techniques that address ambiguity in natural language requirements. Nazir *et al.* [19] shed light on the utilization of NLP in the domain of RE through a systematic literature review. One of the outputs of this research is a collection of leading NLP tools for RE. More recently, Binkhonain and Zhao [20] reviewed machine learning algorithms for identification and classification of non-functional requirements (NFRs), identifying the most popular machine learning algorithms and the most used matrices to measure the performance of these algorithms. Moreover, Dalpiaz *et al.* [4] outline future research directions of NLP in the RE discipline and argue, by means of the article's subtitle, that: "the best is yet to come".

E. RESEARCH GAP

In short, a preliminary review of related literature and trial searches proved that no scientific research has been done on the intersection of GDPR, NLP, and RE (see Section III-B1 for an elaboration on trial searches). One of the studies—the literature study by [19]—aims to investigate the application of NLP techniques in RE and thus partly overlaps with research question 1 of Section III-A1. However, our study presents a more comprehensive approach with a high level of granularity that aims to explore the literature on NLP-based automated GDPR-compliance in RE. This is illustrated by the number of identified studies: [19] arrives—through a systematic literature study—at 27 relevant articles, whereas our mapping study identifies 448 relevant studies. Furthermore, the mentioned study conducts a systematic literature review and not a mapping study, which leaves a gap for our intersectional research. More recently, [21] reported a systematic mapping study of NLP for RE with an ostensibly similar research goal as our first research question. However, closer examination shows that the study in question adopts a more fine-grained approach in classification, whereas we use a higher abstraction level. This is reflected in our data extraction form, where we capture, among other things, research methodology and research theory. Furthermore, our research objective differs from this study as we aim to map the literature to build a bridge that facilitates automated GDPR-compliance in RE, whereas [21] aims to survey the NLP for RE landscape to understand the state-of-the-art and identify open problems. This study was identified later and through different means than the studies reported in the preliminary review of Section II-D and was therefore not included in that Section.

To explore the current state of research as to automated GDPR-compliance using NLP in RE, a systematic study of the literature will be conducted. In general, two types of literature studies can be distinguished: systematic literature reviews and systematic mapping studies. Systematic literature reviews help identify, evaluate, and interpret all

available research relevant to a particular research question, topic area, or phenomenon of interest [22]. On the other hand, systematic mapping studies allow the evidence in a domain to be plotted at a high level of granularity [22], [23].

Due to the novelty of the GDPR and a low number of relevant studies in the preliminary search, we have opted to conduct a systematic mapping study rather than a systematic literature review. Mapping studies provide a structure of the type of research reports and results that have been published by categorizing them, which is often accompanied by a visualization of the results [24]. This approach falls in line with our ambition to map research that centers around automated GDPR-compliance in RE and to identify future research opportunities.

III. MAPPING METHOD

In our systematic mapping process, we followed the guidelines proposed in [22]. This section elaborates on the literature mapping—planning, conducting, and reporting the results—and the rationale behind choices. Planning the mapping process is essential to maintain rigorousness in reviewing and is, for that reason, the starting point of this study.

A. PLANNING THE MAPPING

Mapping protocols are the starting point for every literature mapping and are comprised of the research questions to be addressed through the literature mapping, and a detailed description of the methods that will be used for mapping the literature. By specifying a literature mapping protocol beforehand, research bias possibilities are reduced [22]. In what follows, the research questions, search strategy, study inclusion criteria, and data extraction approach are specified and motivated.

1) RESEARCH QUESTIONS

This study aims to map research conducted on the intersection of GDPR, NLP, and RE. Initial queries showed that research on this topic is scarce. For that reason, we decided to search all possible combinations of the mentioned three domains. The following research questions were formulated.

- **RQ1** "*What NLP approaches are useful for RE and for which RE activity?*" Literature shows that several NLP approaches are available for the different activities in RE. However, no systematic literature mapping exists where these different approaches are mapped with the tasks they can be used for while maintaining the higher goal of automating GDPR-compliance in requirements engineering.
- **RQ2** "*Which NLP approaches are available for achieving GDPR-compliance in organizations?*" In addition to mapping the NLP approaches that were developed specifically for the RE domain, it is also interesting to explore NLP-tasks outside the RE paradigm that focus on achieving GDPR-compliance in organizations. This is especially interesting because of the principle of data protection by design, which implies the need for GDPR-compliant requirements. In other words, we investigate the possibility to learn from these approaches and to transfer this learning to the RE process within software development.
- **RQ3** "*Which state-of-the-art RE solutions are available for achieving GDPR-compliance?*" This question aims to identify state-of-the-art RE solution types that uphold, and possibly facilitate, the GDPR. The answer to this question will provide us with a topical mapping of the RE developments in achieving GDPR-compliance.

B. CONDUCTING THE REVIEW

After drafting the mapping protocol, the actual review can be conducted. This Section describes the search strategy and used data sources, and the study selection process.

1) DATA SOURCES AND SEARCH STRATEGY

The search strategy of a literature mapping is a core component as it affects the number and relevancy of the obtained articles. The approach taken in this research is an automated search through several digital libraries. Part of the search strategy is identifying the appropriate terms that could be used to explore the electronic databases. In this literature mapping study, the main search terms were derived from the three aforementioned research questions. Synonyms of these terms were included by the Boolean OR operator. Finally, these three sets of search terms were concatenated using the Boolean AND operator.

The following digital libraries were searched: ACM Digital Library, Web of Science, IEEE Xplore Digital Library, Scopus, and SpringerLink. ACM Digital Library, IEEE Xplore Digital Library, and SpringerLink are publisher databases, whereas the other databases are indexing services. In concert, these libraries give good coverage of potentially relevant studies.

Furthermore, only studies published between May 2010 and May 2020 were considered due to the novelty and rapid pace of developments in the research domains in question. The initial search query reads as follows: ("**requirements engineering**" *OR* "**requirements analysis**" *OR* "**requirements specification**" *OR* "**requirements elicitation**") *AND* ("**natural language processing**" *OR* **nlp** *OR* "**text mining**") *AND* ("**general data protection regulation**" *OR* **GDPR**).

The results, however, were inadequate. IEEE Xplore Digital Library, for example, returned zero relevant articles on February 11, 2020. This experience led to several trial searches, after which we decided to split the query into three parts, each answering one of the previously mentioned research questions. As a result, the three database-dependent queries (i.e., the search string might have been adapted to satisfy the syntax of the database in question) described in Table 2 were used to search the digital databases.

Search Query 1, consisting of two parts separated by the logical AND operator, aims to identify NLP approaches used

TABLE 2. Search queries used for automated search through electronic databases.

Search Query 1 (SQ1):	(“**requirements engineering**” *OR* “**requirements specification**”) *AND* (“**natural language processing**” *OR* **nlp** *OR* “**text mining**” *OR* “**text analytics**”)
Search Query 2 (SQ2):	(“**natural language processing**” *OR* **nlp** *OR* “**text mining**” *OR* “**text analytics**”) *AND* (“**General Data Protection Regulation**” *OR* **GDPR**)
Search Query 3 (SQ3):	(“**requirements engineering**” *OR* “**requirements specification**”) *AND* (“**General Data Protection Regulation**” *OR* **GDPR**)

TABLE 3. Primary search results.

	Retrieved Studies			
Electronic Database	SQ1	SQ2	SQ3	**Search Fields**
ACM	244	110	16	Full Text
Web of Science	103	4	3	Metadata
IEEE Xplore Digital Library	560	97	46	Full Text & Metadata
Scopus	481	24	27	Metadata
SpringerLink	242	261	101	Full Text
Total	1632	499	190	

for RE. The first part of the query concentrates on NLP, whereas the latter focuses on RE. Search Query 2 aims to capture articles focusing on NLP approaches that facilitate GDPR-compliance. The first part of the query consists of NLP terms, and the second part focuses on the GDPR. Search Query 3 aims to identify RE solutions that support GDPR-guidelines and therefore consists of two parts—the former concentrates on RE and the latter on GDPR. Table 3 shows the primary search outcome, which resulted in a sum of 2321 potentially relevant studies.

2) STUDY SELECTION

After crawling through the digital search space using the queries of Fig. 2 and collecting potential relevant studies, the study selection process was entered. Herein, the potentially relevant studies were subjected to two selection shifts where the studies were assessed for their relevance. However, before this assessment, pre-processing took place where duplicates and outliers (i.e., illegible documents) were filtered out. We used Mendeley, a reference manager system that allows importing BibTeX files and manually mark outliers, for this process. After marking and removing the outliers of the merged data set, the duplicates were removed. The pre-processing and the study selection process is visualized in Fig. 1.

In the first step of the study selection process, studies were held against the inclusion criteria of Table 4. In this step, the title and abstract were reviewed. If the title and abstract were deemed inconclusive—high-quality abstracts are not always guaranteed in IT and software engineering [25]—the article’s Introduction section was reviewed as well.

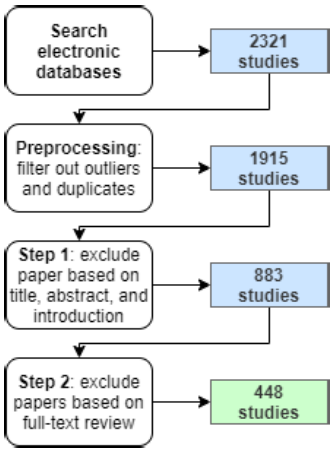

FIGURE 1. The selection procedure of relevant literature.

TABLE 4. Inclusion criteria.

I1.	Peer-reviewed research articles in the form of a journal article, conference paper, symposium article, or book chapter
I2.	Studies focusing on at least one of the aforementioned research questions
I3.	Studies available in full text
I4.	Studies published in English

During this process, we intended to err on the side of cautiousness, that is, if it was not possible to exclude a paper without question, the study was not rejected. In the second stage, the resulting studies were examined in full-text and, again, held against the inclusion criteria. If it was not possible to exclude a paper without question, the study was included. We did not conduct a quality assessment of the selected studies since the focus of this mapping study is to outline all relevant literature, thus focusing on breadth rather than depth [24].

The first inclusion criterion aims to guarantee a level of quality by including solely peer-reviewed research articles. This criterion is based on trial searches and is aimed at excluding, among others, tutorials [26], and study plans [27]. The second inclusion criterion sets the scope for this mapping study by allowing only those articles that are relevant in answering the research questions of this study. For example, some studies propose an NLP tool that itself is compliant with the GDPR, which differs from a approach that aids in achieving GDPR-compliance—a subtle yet significant distinction. Furthermore, only studies available in full-text were considered. Authors of unavailable studies were contacted with the purpose of including their research. Finally, if a study is published other than in the English language, it is excluded. In the next step, the relevant data were extracted from the selected studies.

C. DATA EXTRACTION

A data extraction form was drafted to extract evidence from the reviewed literature. In addition to gathering the data

required for answering the research questions, data extraction forms should collect standard information, including the reviewer's name, date of data extraction, and meta-data of the study in question (e.g., publication details) [22]. Data items I1 through I11 of our data extraction form could be obtained without reading the studies and were therefore considered meta-data. In addition, the data extraction form was composed to capture information related to the research approach and solution of the considered study by capturing the study domain and its corresponding subfield, research method, research theory, research type, knowledge contribution, and solution type. The classifications used for extracting these data are outlined in the next sub-section. The data extraction form itself is presented in Appendix A.

1) CLASSIFICATION

Each reviewed article was classified into its corresponding **study domain**, namely *GDPR, NLP*, or *RE*. These study domains follow naturally from the research questions of this mapping study. Furthermore, these study domains are—as is apparent from the search queries—not mutually exclusive: studies can be classified into multiple domains.

Relative to the study domain, we aimed to capture the **subfield** of the study. The subfield is not a rigid classification and was used to find, for example, activities or concepts that exist within the overarching study domain. To identify the RE activity (or activities) around which the study in review focuses on, we adopted the following classification: *domain understanding & elicitation*, *evaluation & negotiation*, *specification & documentation*, and *quality assurance* [10]. These activities have a logical ordering, however, they are neither sequential, nor linear. Furthermore, the RE activities are iterative in nature. To classify the NLP approaches, we have adopted an a-posteriori approach as will be explained in Section III-D. To determine which part(s) of the GDPR are targeted by the researchers, we use the guiding principles of the GDPR, as outlined in Article 5 [1], as a basis: *lawfulness, fairness and transparency*, *purpose limitation*, *data minimisation*, *accuracy*, *storage limitation*, *integrity and confidentiality*, and *accountability*.

To classify the **research methodology**—or the strategies of inquiry used to answer a specific research question—we distinguish between *quantitative strategies, qualitative strategies, mixed methods*, and *design science methods* [28]. Quantitative strategies comprise research methods characterized by an emphasis on quantitative data. Qualitative strategies, on the other hand, entail research methods that focus more on qualitative data. However, some strategies encompass both quantitative and qualitative-oriented research methods; these strategies are labeled as mixed methods. Finally, we also consider design science methods, which are described as strategies that focus on building and evaluating new and innovative artifacts. Design science is solution-oriented and therefore widely interpretable. For that reason, we will label articles as upholding design science methodologies based on both explicit and implicit indicators rather than fixating on a verbatim mention in the paper of a particular design science method. For example, we perceive [29] as following a design science method since it presents an innovative artifact in the form of a tool for automated ambiguity detection, even though the study does not explicitly mention design science as research methodology. Considering the nature of our research questions, which aim to collect artifacts (e.g., approaches and solutions), we assume that a significant number of studies will utilize design science methodologies. For that reason, we decided to use existing classifications that focus on design science as a jumping-off point.

In terms of classifying the **research theory**—a system of constructs and corresponding relationships that collectively present a logical, systematic, and coherent explanation of a phenomenon of interest—that results from the study under review, we differentiate between the following theory types: *analysis, explanation, prediction, explanation and prediction, design and action* [28]. Analysis theories focus on describing phenomena by stating *what* is. Explanation theories supplement this by focusing also on describing *how*, *why*, *when*, and *where* phenomena occur. Prediction theories focus on describing *what* is and *what will* be, without regarding the cause. Explanation and prediction theories, combined, attempt to merge the previous two foci, meaning, they focus on predicting what will be and its cause to be. Finally, design and action theories concentrate on *how* to do something. Similar to our research methodology classification, this classification was not solely based on explicit indicators: implicit indicators were considered as well. Reference [29] for example, provides prescriptions for constructing an artifact and is therefore considered to contribute a design and action theory—without explicitly mentioning this contribution.

To classify the **research type** of the papers under review, we use the classification of [30]—*validation research, evaluation research, solution proposal, philosophical papers, opinion papers*, and *experience papers*—as a starting point. Since this classification is composed of a design science lens, we have gradually, during the review process, added research types that we deemed necessary for classifying papers that uphold different research methods than design science. Similar to the study domain, studies can be categorized into multiple research types. For instance, [31] proposes a method (i.e., solution proposal) and conducts a controlled experiment (i.e., validation research).

As to the **knowledge contribution** of the studies that uphold—whether this methodology is explicitly mentioned or not—a design science research method, we have adopted the design science contribution framework as described in [32]. This quadrantal framework consists of improvement (new solutions for known problems), invention (new solutions for new problems), routine design (known solutions for known problems), and exaptation (known solutions for new problems). On a par with our classification philosophy of research methodology and research theory, we based our classification of knowledge contribution on both implicit as explicit indicators. Again, we will

consider [29] to elucidate our approach. The referenced study aims to automate the practice of ambiguity detection by introducing an NLP-based tool. We consider this to be an implicit indication of a so-called improvement, for the study introduces a new solution for the known problem of ambiguity detection.

Finally, we captured the **solution type**, or the manifestation of this contributed knowledge, in one of the following categories as used in [24]: *metric*, *tool*, *model*, *method*, and *process*. Consistent with our approach for research type classification, we have, during the course of reviewing, added missing solution types. Furthermore, studies can propose different solution types—for example, a method and a methodology [33]. The latter is also an example of a solution type that was added during the reviewing process.

D. DATA SYNTHESIS

The synthesis of the collected data was done by counting the number of studies classified in each of the defined data items of the data extraction form. The data item of Author's affiliation (I7) requires more clarification. For each study, we captured the country or region of affiliation of the corresponding author. Each study could be linked to multiple countries or regions; however, we did not assign the same country multiple times to a study.

The answer to RQ1 necessitated more analysis since the collected NLP approaches were not directly classified during the review process. As outlined in the previous subsection, the classification for this particular data item was derived by reasoning from the results. We did not attempt to superimpose a classification, rather, we decided to capture NLP approaches as mentioned in the corresponding studies. As a result, the gathered approaches may slightly differ in granularity. For example, the tasks of tokenizing and pre-processing are mentioned frequently in the set of retrieved studies. The latter could be interpreted as an abstraction or generalisation of the former. However, we decided not to group them because tokenizing was mentioned more frequently. Nevertheless, we decided to arrange frequent co-occurring text cleaning activities such as punctuation removal, lowercasing, and stopword removal under the label of ''pre-processing''. Another example is that we categorized different classification algorithms (e.g., Naive Bayes Classification) under one label of text-classification.

As for the GDPR related studies, we decided to note—alongside the GDPR concepts as per the data extraction form—the relevant GDPR article, if applicable. For example, if a study focuses on the concept of Consent, we annotate the relevant GDPR article 7.

Finally, as explained in Section III-C1, studies may be allocated to multiple study domains, RE activities, research types, and solution types. As a result, it may be that, in the visualizations of the next Section, the same study is counted more than once—for example, when a study is allocated to multiple RE activities (see Fig. 5). However, these counts are accompanied by their relative frequency, calculated based on the unique sum of studies.

IV. RESULTS

This section describes the results of the mapping study. First, we present an outline of the data through the synthesis of the metadata. Consequently, we present the main contributions of this research by answering the research questions of this study. The final list of retrieved studies and outcome of this systematic mapping study can be found at the following repository: https://aberkane.github.io/SMS_GDPR-NLP-RE.

A. STUDY SELECTION RESULTS

The study selection process resulted in 448 relevant studies. The majority of these studies—420 studies—were allocated to the study domain of RQ1, which centers around NLP and RE. The questioning of RQ2, focusing on GDPR and NLP resulted in nine studies. RQ3, which centers around GDPR and RE, resulted in 20 studies. One of the retrieved studies fell within all pairwise research field combinations (i.e., NLP & RE, NLP & GDPR, and GDPR & RE) [34]. The referenced study presents EPICUREAN, a recommender-based privacy requirements elicitation approach (EPICUREAN) which uses NLP and machine learning techniques in the RE activity of domain understanding & elicitation to determine and recommend appropriate privacy settings to the user and hereby simplifying privacy settings concerning the GDPR.

The lion's share of the retrieved studies was authored by researchers solely active in academia (88%). A smaller part—14 studies—were undertaken by authors affiliated with the industry. There were also collaborations between different backgrounds: 39 studies were conducted by both authors affiliated with academia and industry, and two of the studies had authors affiliated with both academia and governmental organizations.

TABLE 5. Distribution of the retrieved studies over electronic databases and the total number of retrieved studies in unique terms.

	Electronic Database					
Research Questions	ACM	Web of Science	IEEE Xplore	Scopus	Springer-Link	**Total**
RQ1	71	66	166	252	73	**419**
RQ2	2	0	1	3	4	**9**
RQ3	2	2	6	12	9	**20**

Regarding the electronic database from which the studies were retrieved, we observe that 267 studies (59,6%) were found in Scopus, 173 studies were retrieved from IEEE Xplore Digital Library, 86 studies were retrieved from SpringerLink, 75 studies were retrieved from ACM Digital Library, and 69 studies were retrieved from the Web of Science database. A summary is given in Table 5. Overlap between the different databases did occur.

Fig. 2 presents the temporal evolution of the number of studies, distinguishing between the different combinations. An upward trend of research related to NLP & RE can be observed—especially after 2014—with an optimum noticeable in 2019. It seems that the requirements engineering

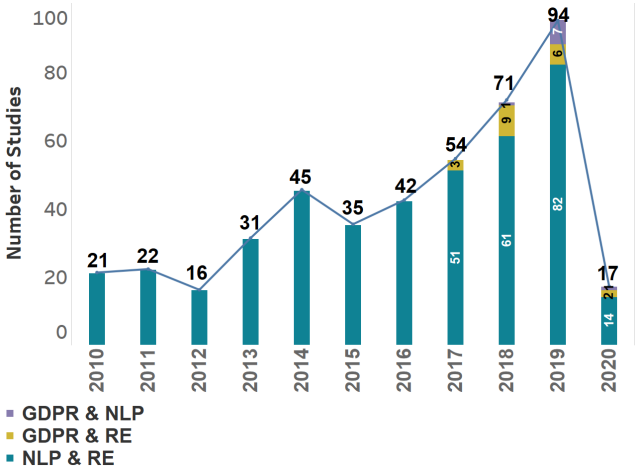

FIGURE 2. Number of studies per year (2010-2020).

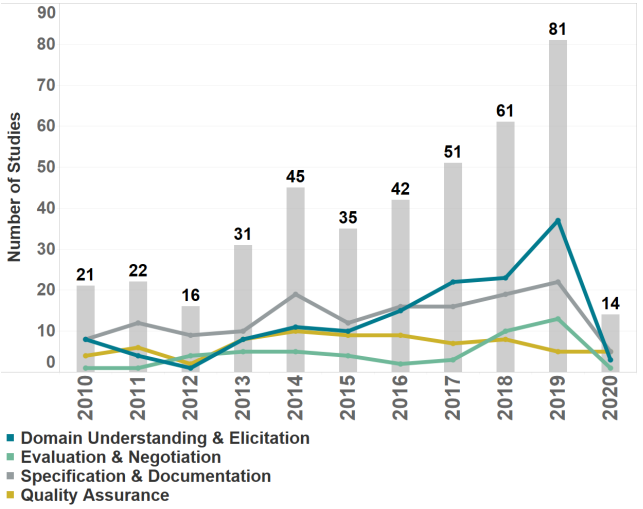

FIGURE 3. Publication trend of studies related to RQ1.

research community is gravitating towards the use of NLP for RE practice. As for the research related to GDPR, it is not unexpected that the number of studies related to GDPR is small, since the GDPR was only adopted in May 2018 [35]. From 2018 onwards, an increment is visible. This rise commences with research related to GDPR and RE, and in the subsequent year, the emergence of GDPR research related to NLP is visible. However, this period is too short to draw conclusions, and it remains to be seen whether this trend will continue in the future. Note that in 2019, one study was labeled as related to both NLP & RE, and GDPR & RE.

B. RQ1. WHAT NLP APPROACHES ARE USEFUL FOR RE AND FOR WHICH RE ACTIVITY?

The first research question aims to identify NLP approaches which are useful for RE and, in particular, for which RE activities. In this mapping study, 420 studies were identified as relevant to this research question, i.e., focusing on both NLP and RE. The details of the mapping can be found in the aforementioned repository.

The majority of the retrieved studies relevant to RQ1, namely 370, were conducted by authors exclusively affiliated with academia (88.1%), 34 of the studies had authors affiliated with both academia and industry (8.1%), 13 studies were conducted by researchers uniquely affiliated with the industry (3.1%), and two studies were conducted by researchers affiliated with academia and governmental organizations (0.5%).

Fig. 3 depicts the studies' publication trend while distinguishing between the different RE activities. As stated before, 420 studies were identified as relevant to RQ1. It follows—while being cognizant of the fact that not the whole of 2020 was taken into consideration—that the average publication number is 42 studies per year. From 2016 onward, studies related to all RE activities, except quality assurance, show an increasing trend with both an individual and a collective optimum in 2019. The retrieved studies were primarily centered around the RE activities of domain understanding & elicitation (143) and specification & documentation (148). The activities of evaluation & negotiation and quality assurance—accounting 49 and 73 studies, respectively—seem to have received less attention, at least in comparison to the previously mentioned activities, by researchers and this is reflected throughout this section. Finally, nine of the studies were not classified to a specific RE activity.

Fig. 4 shows the geographical distribution of the studies related to RQ1. The list is led by the USA, Germany, and India, each associated with 15%, 13.1%, 12.9% of the studies. They are followed—at appropriate distance—by Italy, China, and the UK, jointly affiliated with 21.2% of the studies. After that, we find Japan, Canada, the Netherlands, Pakistan, Malaysia, Luxembourg, and France—cumulatively related to 23.3% of the retrieved studies. What remains is a list of countries affiliated with less than ten studies each.

Concerning the research methodology, 91.4% of the studies were classified as using design science research methods, 3.3% of the studies used quantitative research methods, 5% used qualitative research methods, and 0.2% of the studies used mixed research methods (i.e., both qualitative and quantitative research methods). As argued in Section III-C1, the nature of our research question tends to identify studies engaging in design science, hence, this result was to be expected. The frequency of the research methodology as identified in the retrieved studies related to NLP and RE are presented in Fig. 5. Moreover, Fig. 6 shows a detailed analysis of the research methodology and the corresponding research theory. Since the vast majority of studies focuses on design science research methods, it is in line with expectations that a significant amount of studies will adopt a design and science theory to instruct how to design an artifact. Furthermore, regarding the studies where knowledge contribution was relevant, 359 studies were considered to contribute in the form of an improvement—of which 354 related to design and research theory—whereas one study was classified as

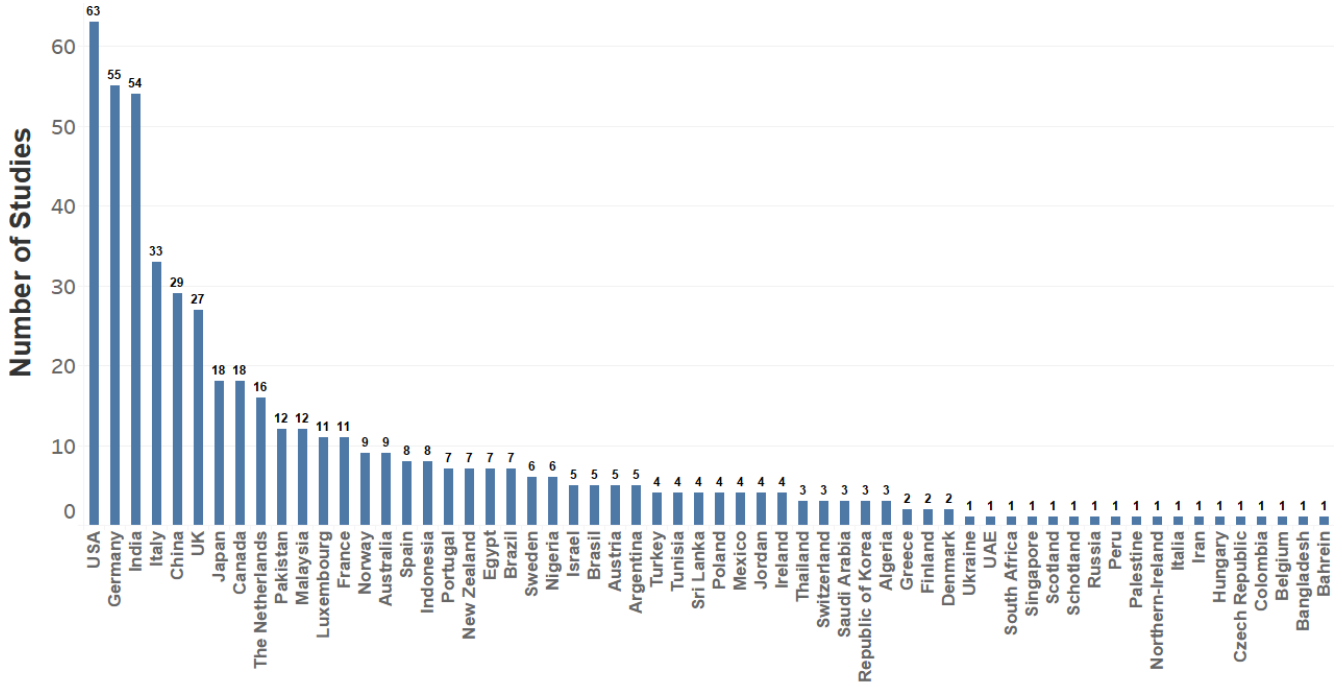

FIGURE 4. Geographical distribution of studies related to RQ1.

Research Methodology	Domain Understanding & Elicitation	Evaluation & Negotiation	Specification & Documentation	Quality Assurance	Not Specified
design science methods	127 (30,2%)	46 (11,0%)	142 (33,8%)	65 (15,5%)	6 (1,4%)
mixed methods		1 (0,2%)			
qualitative strategies	7 (1,7%)	1 (0,2%)	5 (1,2%)	5 (1,2%)	3 (0,7%)
quantitative strategies	9 (2,1%)	1 (0,2%)	1 (0,2%)	3 (0,7%)	

FIGURE 5. Research methodology and corresponding RE activity in absolute (and relative) terms as retrieved from studies related to RQ1.

Research Theory	design science methods	qualitative strategies	quantitative strategies	mixed methods
analysis	11 (2,6%)	20 (4,8%)	10 (2,4%)	
design and action	366 (87,6%)		1 (0,2%)	1 (0,2%)
explanation		1 (0,2%)	2 (0,5%)	
explanation and prediction	6 (1,4%)			
prediction	1 (0,2%)		1 (0,2%)	

FIGURE 6. Research theory and research methodology in absolute (and relative) terms as retrieved from studies related to RQ1.

Research Type	Domain Understanding & Elicitation	Evaluation & Negotiation	Specification & Documentation	Quality Assurance	Not Specified
evaluation research	9 (2,1%)	6 (1,4%)	2 (0,5%)	2 (0,5%)	
experience paper	4 (1,0%)	2 (0,5%)		2 (0,5%)	4 (1,0%)
literature review	8 (1,9%)		5 (1,2%)	3 (0,7%)	2 (0,5%)
philosophical paper					1 (0,2%)
position paper	1 (0,2%)	1 (0,2%)		1 (0,2%)	
solution proposal	112 (26,7%)	38 (9,0%)	134 (31,9%)	58 (13,8%)	1 (0,2%)
validation research	12 (2,9%)	6 (1,4%)	10 (2,4%)	8 (1,9%)	

FIGURE 7. Research types and RE activities in absolute (and relative) terms as retrieved from studies related to RQ1.

routine design. Improvement is essentially, the underlying idea of incorporating NLP approaches that are new to the RE field, to solve existing RE problems, for improvement means developing new solutions for known problems [32].

Initially, as described in Section III-C1, we distinguished between 5 different research types. However, during the mapping process, we identified the need to add one research type: literature reviews. Fig. 7 summarizes the frequency of identified research types combined with the related RE activity. As expected and consistent with our previous analysis, solution type was the most predominant research type as it follows from the design science methodology and the design and action theory type. However, this disproportion can be worrying since different research types are important to enrich the scientific body of literature, for example, by providing new insights through philosophical papers. Furthermore, solution type was mainly concentrated in the RE activities of domain understanding & elicitation and evaluation & negotiation.

Fig. 8 lists the identified solution types and their frequency, in both absolute and relative terms, related to the RE activity upon which the corresponding study focuses. As explained in Section III-C1, we did not maintain a rigid approach towards the different solution types. As a result,

Solution Type	Domain Understanding & Elicitation	Evaluation & Negotiation	Specification & Documentation	Quality Assurance	Not Specified
algorithm	1 (0,2%)	1 (0,2%)	7 (1,7%)	2 (0,5%)	
approach	31 (7,4%)	10 (2,4%)	44 (10,5%)	16 (3,8%)	1 (0,2%)
architecture	1 (0,2%)	2 (0,5%)	2 (0,5%)	1 (0,2%)	
corpus			1 (0,2%)		
dataset				1 (0,2%)	
framework	13 (3,1%)	2 (0,5%)	13 (3,1%)	9 (2,1%)	1 (0,2%)
gold standard	1 (0,2%)		1 (0,2%)		
guidelines			1 (0,2%)		
lexicon	1 (0,2%)				
method	16 (3,8%)	5 (1,2%)	19 (4,5%)	11 (2,6%)	
methodology	5 (1,2%)	1 (0,2%)	3 (0,7%)		
metrics	1 (0,2%)		2 (0,5%)		
model	5 (1,2%)	3 (0,7%)	2 (0,5%)	1 (0,2%)	
pattern			1 (0,2%)		
pipeline	1 (0,2%)	1 (0,2%)			
principles	1 (0,2%)				
process	5 (1,2%)		1 (0,2%)	1 (0,2%)	
program	1 (0,2%)				
research-line	1 (0,2%)	1 (0,2%)			
scheme				1 (0,2%)	
state-of-the-art				1 (0,2%)	
system	9 (2,1%)		6 (1,4%)		
taxonomy	1 (0,2%)				
technique	9 (2,1%)	4 (1,0%)	7 (1,7%)	1 (0,2%)	
tool	30 (7,1%)	14 (3,3%)	38 (9,0%)	20 (4,8%)	

FIGURE 8. Solution types and RE activities in absolute (and relative) terms as retrieved from studies related to RQ1.

NLP Approach	Domain Understanding & Elicitation	Evaluation & Negotiation	Specification & Documentation	Quality Assurance	Not Specified
dependency text-parsing	2 (0,5%)	5 (1,2%)	12 (2,9%)	3 (0,7%)	
feature extraction	12 (2,9%)	4 (1,0%)	7 (1,7%)	4 (1,0%)	2 (0,5%)
lemmatization	25 (6,0%)	6 (1,4%)	13 (3,1%)	10 (2,4%)	1 (0,2%)
n-grams	14 (3,3%)	3 (0,7%)	3 (0,7%)	2 (0,5%)	1 (0,2%)
POS-tagging	51 (12,1%)	23 (5,7%)	64 (15,2%)	26 (6,2%)	2 (0,5%)
pre-processing	49 (11,7%)	13 (3,1%)	35 (8,3%)	21 (5,0%)	1 (0,2%)
semantic analysis	4 (1,0%)	6 (1,4%)	10 (2,4%)	1 (0,2%)	
sentence parsing	8 (1,9%)	3 (0,7%)	17 (4,0%)	6 (1,4%)	
stemming	15 (3,6%)	8 (1,9%)	16 (3,8%)	7 (1,7%)	
text-chunking	2 (0,5%)	2 (0,5%)	18 (4,3%)	3 (0,7%)	2 (0,5%)
text-classification	48 (11,7%)	12 (3,1%)	27 (6,4%)	13 (3,1%)	5 (1,2%)
text-clustering	15 (3,6%)	2 (0,7%)	5 (1,2%)	5 (1,2%)	2 (0,5%)
text-parsing	13 (3,1%)	3 (0,7%)	34 (8,3%)	17 (4,0%)	1 (0,2%)
TF-IDF-weighting	13 (3,1%)	5 (1,2%)	3 (0,7%)	1 (0,2%)	1 (0,2%)
tokenizing	24 (5,7%)	17 (4,0%)	39 (9,3%)	23 (5,5%)	2 (0,5%)

FIGURE 9. 15 Most frequently occurring NLP approaches in absolute (and relative) terms as retrieved from studies related to RQ1 against the corresponding RE activity.

during the review process, the initial taxonomy of solution types was expanded to 25. The majority of the proposed solutions were concentrated—in line with the analysis in this Section hitherto—in the RE activities of domain understanding & elicitation and specification & documentation. The most frequently proposed solution is a Tool (24.1%) closely followed by an Approach (23.9%).

Research question 1 aimed to identify useful NLP approaches for RE. Fig. 9 lists the most relevant information related to this research question; namely, the 15 most frequently occurring NLP approaches allocated to the different RE activities. In total, 199 approaches were identified. The NLP approaches of part-of-speech (POS) tagging, preprocessing, and text-classification lead the way with 165, 117, and 104 occurrences, respectively. The complete list can be found in the previously mentioned repository.

In Table 6, the most popular conferences, workshops, and symposia are listed alongside the respective numbers of publications. Moreover, Table 7 depicts the most popular journals and their respective number of publications. Finally, in Table 8 we present the most popular studies in terms of their number of citations according to Google Scholar.

TABLE 6. Most popular conferences, workshops and symposia.

Venue	Number of Publications
International Requirements Engineering Conference (RE)	49
International Conference on Requirements Engineering: Foundation for Software Quality (REFSQ)	34
Requirements Engineering Conference Workshops (REW)	16
Artificial Intelligence for Requirements Engineering (AIRE)	9
Automated Software Engineering (ASE)	7
International Symposium on Empirical Software Engineering and Measurement (ESEM)	7
Innovations on Software Engineering Conference (ISEC)	7
Symposium on Applied Computing (SAC)	7
Conference on Advanced Information Systems Engineering (CAiSE)	6
International Conference on Applications of Natural Language to Information Systems (APSEC)	5

C. RQ2. WHICH NLP APPROACHES ARE AVAILABLE FOR ACHIEVING GDPR-COMPLIANCE IN ORGANIZATIONS?

The second research question aimed to identify studies that mention NLP approaches used to achieve GDPR-compliance in organizations. Table 10 in Appendix B lists the nine identified studies alongside their publication year, venue, the number of citations, identified NLP task, identified GDPR concept and—if applicable—its corresponding GDPR article, and the proposed solution type. All studies were published as from 2018. The complete mapping can be found in the previously mentioned repository.

As presented in Table 10, the most frequently occurring NLP approach is text-classification (4), followed by preprocessing (2). Moreover, all of the identified studies related

TABLE 7. Most popular journals.

Journal title	Number of Publications
Requirements Engineering	14
IEEE Transactions on Software Engineering	7
Automated Software Engineering	4
Empirical Software Engineering	4
Information and Software Technology	4
ACM Transactions on Software Engineering and Methodology	3
Journal of Systems and Software	3
Software Quality Journal	3
ACM SIGSOFT Software Engineering Notes	2
IEEE Software	2

TABLE 8. Most popular studies.

Study	Title	Number of Citations
[36]	How do users like this feature? A fine grained sentiment analysis of App reviews	451
[37]	Bug report, feature request, or simply praise? On automatically classifying app reviews	308
[38]	On the automatic classification of app reviews	133
[39]	Improving agile requirements: the Quality User Story framework and tool	117
[40]	Analysing anaphoric ambiguity in natural language requirements	107
[41]	Supporting domain analysis through mining and recommending features from online product listings	106
[42]	Automated extraction of security policies from natural-language software documents	105
[43]	Semi-automatic generation of UML models from natural language requirements	105
[44]	Automated checking of conformance to requirements templates using natural language processing	102
[45]	Empirical principles and an industrial case study in retrieving equivalent requirements via natural language processing techniques	96

to RQ2 used design science research methods as a research methodology. Also, each study was allocated to upholding a design and action research theory. As for the research type, seven out of nine studies were categorized as proposing a solution, of which one study conducted a validation research as well. The remaining two other studies were categorized as evaluation research and validation research. Regarding the knowledge contribution, eight out of nine studies were labeled as contributing through improvement, while one study was labeled as proposing a new solution for a new problem (i.e. invention).

Fig. 10 details the most frequent GDPR concepts related to RQ2. Four different concepts were identified in the

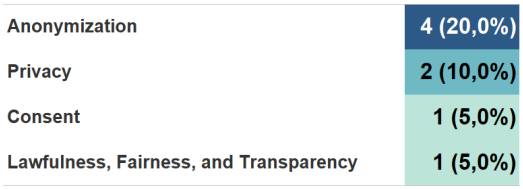

FIGURE 10. Most frequent GDPR concepts in absolute (and relative) terms as retrieved from studies related to RQ2.

nine studies. Anonymization was identified four times, Privacy twice, Consent once, and Lawfulness, Fairness and Transparency once as well.

D. RQ3. WHICH STATE-OF-THE-ART RE SOLUTIONS ARE AVAILABLE FOR ACHIEVING GDPR-COMPLIANCE?

This mapping study identified 20 studies that discussed RE solutions for achieving GDPR-compliance in organizations. Fig. 11 depicts the distribution of the number of studies to the corresponding RE activities. Again, research was mainly centered around the RE activities of domain understanding & elicitation and evaluation & negotiation, accounting for ten and six studies, respectively. Four studies were allocated to evaluation & negotiation, one study was allocated to the activity of quality assurance, and, finally, 1 study was not assigned.

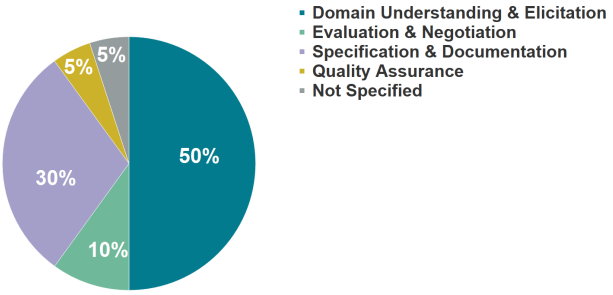

FIGURE 11. Identified RE activities in studies related to RQ3.

All of the retrieved studies were labeled as adopting a design science methodology and contributing a design and action research theory, which follows the trend of studies related to RQ2. Regarding the research type, two studies were classified as position papers [46], [47], whereas the rest were classified as solution proposals. This emphasis on solution proposals is no different than what we have observed in the studies related to the previous research questions.

The complete list of studies that were collected to answer RQ3 is presented in Table 11 of Appendix B. Each study is described by its reference number, year of publication, RE activity, number of citations, publication type, and—the key to RQ3—its proposed solution type. The full mapping is available at the previously mentioned repository. Five different solution types were observed. The majority of the retrieved studies proposed Methods (4) and Approaches (4). Additionally, 35% of the studies related to RQ3 proposed

Frameworks (3), Architectures (2), or Tools. All studies were published after 2016, which is the year in which the GDPR was adopted.

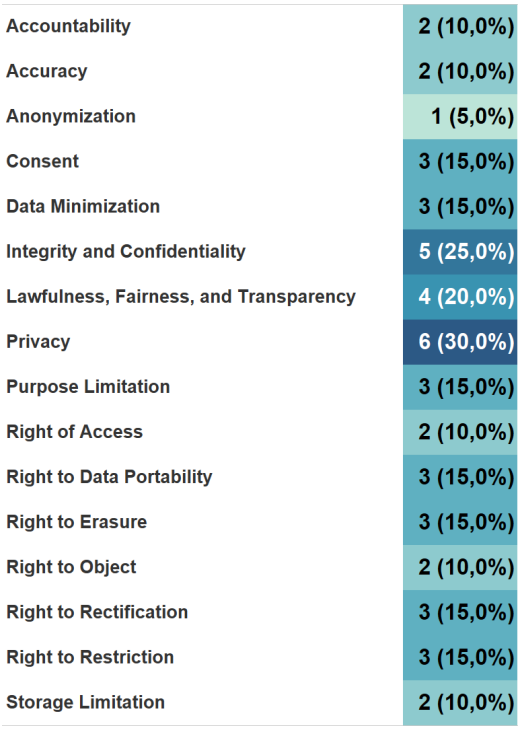

FIGURE 12. Most popular GDPR concepts identified in the retrieved studies related to RQ3.

In Fig. 12, the most popular GDPR concepts according to the retrieved studies are presented in a tree-map with their corresponding share expressed in terms of percentages. Studies that did not specify a specific or limited set of concept (i.e., they discussed the GDPR in general), were not considered in this analysis.

The concepts of Integrity and Confidentiality—as described in Article 5 of the GDPR [1]—and Privacy were most prevalent, accounting for 25% and 30% of the retrieved studies, respectively. The concepts of Privacy and—as described in Article 5 of the GDPR [1]—Integrity and Confidentiality, and Lawfulness, Fairness and Transparency were most prevalent, accounting for 30%, 25%, and 20% of the retrieved studies, respectively. Next, seven concepts were identified, each present in 15% of the retrieved studies related to RQ3: Consent (Article 7), Data Minimization (Article 5), Purpose Limitation (Article 5), Right to Data Portability (Article 20), Right to Erasure or ''Right to be Forgotten'' (Article 17), Right to Rectification (Article 18 & 19), and Right to Restriction (Article 18). Five concepts were identified with each concept present in 10% of the studies: Accountability (Article 5), Accuracy (Article 5), Right of Access (Article 15), Right to Object (Article 21), and Storage Limitation (Article 5). Finally, the concept of Anonymization occurred once.

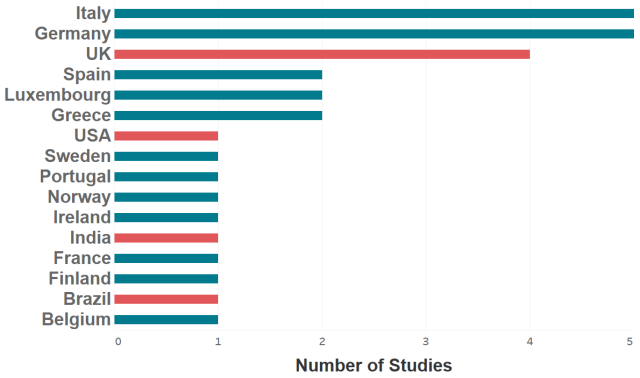

FIGURE 13. Geographical distribution of studies in GDPR and requirements engineering based on the corresponding researchers: the blue-colored bars indicate presence in the EEA, whereas the red-colored bars indicate absence in the EEA.

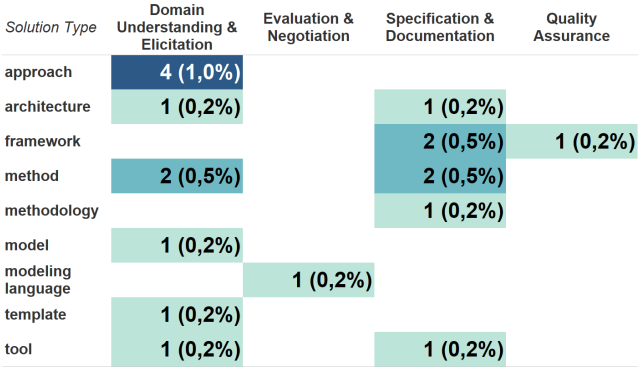

FIGURE 14. Most frequent solution types in absolute (and relative) terms as retrieved from studies related to RQ3.

As outlined in Section II-A, the GDPR applies to all countries in the EEA. In fact, when personal data is transferred outside the EEA, it remains under the safeguards of the GDPR. Fig. 13 details the geographical distribution of retrieved studies related to RQ3 based on researchers' affiliated countries. Studies can be linked to different countries based on the corresponding researchers' country affiliation. However, a study can only be linked once to the same country, despite the number of authors. The countries are divided based on their presence in the EEA: the blue-colored bars highlights a country's presence in the EEA, whereas a red color highlights its absence. A significant part of research concerning GDPR and RE was conducted by researchers affiliated with Italy (5 studies) and Germany (5 studies). Furthermore, four of the studies were carried out by researchers affiliated with the United Kingdom, which is regulated by the United Kingdom Data Protection Regulation (UK-GDPR) as a result of the British exit from the European Union [48]. The remainder of the affiliated countries consists of countries within the EEA, except for the USA, India, and Brazil.

As described at the onset of this Section, all retrieved studies relevant to RQ3 adopted a design science methodology. Consequently, the majority of the retrieved studies proposed a solution to a real-world problem. Fig. 14 sets the retrieved

solutions against their frequency in the different RE activities. The leading solutions were Approach and Method, proposed as a solution by four studies each. Every Approach was proposed from the perspective of domain understanding & elicitation, whereas the RE standpoints of the Methods were equally divided between domain understanding & elicitation and evaluation & negotiation. The solution type of Framework was put forward thrice: twice from the evaluation & negotiation viewpoint and once from quality assurance. Next, solutions were proposed in the form of an Architecture and Tool—twice each. These solutions were proposed for the RE activities of domain understanding & elicitation and Specification, and quality assurance—equally distributed. Moreover, four solutions were proposed in the form of a Model, Template, Modeling Language, and Methodology.

V. DISCUSSION

In this systematic mapping study, we have systematically explored literature where the fields of GDPR, NLP, and RE coincide. In this section, we will present the main findings and identified gaps, and discuss the limitations of this research.

A. MAIN FINDINGS

The first research question aimed to map NLP approaches to the corresponding RE activities in which they were used. To address this first research question, we have identified 420 studies and used a data extraction form to capture relevant data. The publication trend shows a clear rise of publications in the last ten years with an optimum in 2019. We have collected 199 different NLP approaches, alongside the RE activity in which they were used. A summary of the results is presented in Fig. 9.

The second research question focused on mapping NLP approaches useful for GDPR-compliance in organizations. Not much can be said regarding the publication trend, as it awaits to be seen how this recent research stream will develop. Nevertheless, we identified nine relevant studies (see Table 10), comprising 17 different NLP techniques. Furthermore, the studies discussed four different GDPR concepts. The NLP techniques are listed in Table 10 and the GDPR concepts are presented in Fig. 10.

The third research question intended to explore the different types of RE solutions for achieving GDPR-compliance in RE. Similar to the previous research question, no profound conclusion can be made from the results as this research stream is still in its infancy. We identified 20 studies from which nine different solution types were extracted as listed in Fig. 14. The list of retrieved studies with a set number of data elements are listed in Table 11.

Next to the key findings that addressed our research questions, we discovered several interesting phenomena. First, as can be observed from the data, the mapping revealed a tendency among researchers from the RE community to focus mainly on the RE activities of domain understanding & elicitation and specification & documentation, while less attention is paid to the activities of evaluation & negotiation and quality assurance. This was true for research related to both RQ1 and RQ2. For studies related to RQ1, this may imply that researchers overlooked the possibility of using NLP for the activities of evaluation & negotiation, and quality assurance. Another possibility is that RE tasks in the latter are less suited for automating with NLP.

Second, it could be anticipated from the line of questioning in our research questions that the resulting studies will gravitate towards design science research. Therefore, it is not unexpected that the vast majority of studies focused on proposing a solution—as solution proposals are the crux of design science. However, this is done at the expense of other research types and has, consequently, a detrimental effect on the diversity of a research domain.

1) BRIDGING THE GAP

As mentioned at the beginning of this Section, this mapping study shows that only one study by Stach and Steimle [34] emerged on the convergence of GDPR, NLP, and RE. Despite this scarcity, we have—during the review process—identified possibilities for bridging these research fields to achieve data protection by design in RE and thus nullifying non-compliance during software development. In the next paragraphs, we will outline some of the potentials that have risen from examining research related to RQ2 and RQ3. We focus primarily on literature related to these two questions because they addressed the GDPR explicitly. By scrutinizing the solutions proposed by literature related to NLP and GDPR (RQ2), we can reflect on utilizing these NLP-based solutions for RE problems. On the other hand, problems addressed by literature related to GDPR and RE (RQ3) may provide opportunities for NLP to, for example, automate (parts of) the proposed solutions.

Examining the proposed approaches in the retrieved studies related to RQ2 gives rise to the opportunity of introducing NLP-based machine learning techniques to classification problems in the domain of GDPR and RE. Several studies related to RQ2 explore the concept of personal data identification and anonymization in documents using NLP-based machine learning techniques to comply with the GDPR [49], [50]. We argue that this approach can be introduced, albeit with few changes, to the RE domain; particularly for requirements written in natural language. This approach can be used, for instance, to validate the requirements documents, produced during the specification & documentation activity [10], on GDPR-compliance. However, since it is unlikely that these documents contains personal data, it is more interesting to assess whether the documents comply with other GDPR-requirements; for instance, assessing whether the data that the system-to-be will process is collected ''in a form which permits identification of data subjects for no longer than is necessary'' [1].

From the same pool of literature, that is literature related to RQ2, the opportunity of using NLP-based machine learning techniques arises also from the study by Chang *et al.* [51]

TABLE 9. Data extraction form.

Identifier	Data Item	Description
I1	Reviewer	*the name of the reviewer*
I2	Extraction date	*the date of data extraction*
I3	Publication year	*the publication year of the study*
I4	Title	*the title of the study*
I5	Venue	*the name of the publication venue*
I6	Author(s)	*the name(s) of the author(s)*
I7	Authors' affiliation	*academia, industry, government organisation, or not clear*
I8	Region	*country or region of affiliation*
I9	Database	*the database from which the study was retrieved*
I10	Citations	*the number of citations according to Google Scholar*
I11	Publication type	*journal article, conference paper, workshop paper, symposium paper, or book chapter*
I12	Study domain	*requirements engineering, natural language processing, and/or GDPR*
I13	↪ Subfield	*RE activity, NLP approach, or GDPR concept*
I14	Research methodology	*design science methods, quantitative strategies, qualitative strategies, or mixed methods*
I15	Research theory	*analysis, explanation, prediction, explanation & prediction, or design & action*
I16	Research type	*validation research, evaluation research, solution proposal, philosophical paper, opinion paper, experience paper, or literature review*
I17	Knowledge contribution	*improvement, invention, routine design, or exaptation*
I18	Solution type	*e.g., metric, tool, model, method, framework*

TABLE 10. List of retrieved studies centered around NLP and GDPR.

Study	Year	Venue	Number of Citations	NLP Approach	GDPR Concept	Relevant GDPR Article	Solution
[55]	2019	KDD *(conference)*	23	text-generation; text-classification; word frequency	Consent; Lawfulness, Fairness, and Transparency	5	Method
[50]	2018	ADBIS *(conference)*	3	text-classification; regular expressions; named-entity recognition	Anonymization	5, 7	System
[51]	2019	WASA *(conference)*	2	text-classification; tokenizing	Privacy		System
[56]	2019	ICEIS *(conference)*	1	entity recognition	Anonymization		Architecture
[57]	2019	DDP *(conference)*	0	pre-processing; named-entity recognition			Algorithm
[58]	2019	LOD *(conference)*	0	TF-IDF-weighting; skip-thought vectors; text-classification	Anonymization		Algorithm
[59]	2019	PRO-VE *(conference)*	0	pre-processing; text-classification	Anonymization		Method
[49]	2019	IBIMA *(conference)*	0	tokenizing; text segmentation; POS-tagging; text-classification; named-entity recognition; disambiguation; stemming; lemmatization; n-grams	Privacy		Approach
[60]	2020	SAC *(symposium)*	0	text-classification; convolutional neural network			Tool

where the privacy concerns of mobile application users' are identified and held against the privacy policy in practice. When introduced to the RE discipline, this assessment may lead to new requirements that address the raised privacy concerns and help achieve data protection by design by meeting the transparency requirement of the GDPR.

Another possibility for NLP-based machine learning in GDPR-related research lies in Crowd-based RE. The study by Groen and Ochs [52], identified through RQ3, discusses the possibility of online user feedback being subject to the GDPR. Furthermore, the authors suggest anonymization as a viable solution. Achieving anonymization can be recognized as a NLP-based classification problem (i.e., classifying whether feedback contains personal data), thus indicating bridging possibilities between GDPR and RE on the one hand, and NLP on the other hand.

Finally, literature related to the GDPR and RE (RQ3) charts possibilities of utilizing NLP to assist with manual tasks. The literature discusses, among other things, methods and approaches towards achieving GDPR-compliance during the requirements engineering process, which include the manual

TABLE 11. List of retrieved studies that focus on GDPR and RE.

Study	Year	Venue	Number of Citations	RE Stage	GDPR Concept	Relevant GDPR Article	Solution
[14]	2018	RE *(conference)*	30	Domain understanding & elicitation	Lawfulness, Fairness, and Transparency; Purpose Limitation; Data Minimization; Accuracy; Storage Limitation; Integrity and Confidentiality; Accountability	5	Approach
[46]	2018	EuroS&PW *(symposium)*	25	Not specified			
[61]	2017	PoEM *(conference)*	21	Specification & documentation	Privacy		Method
[62]	2017	APF *(conference)*	16	Quality assurance	Privacy		Framework
[63]	2017	Information (*Journal*)	15	Domain understanding & elicitation	Consent; Individual Participation and Access; Lawfulness, Fairness, and Transparency; Right of Access; Right to Rectification; Right to Erasure; Right to Data Portability; Right to Restriction	5, 7, 15, 17, 18, 19, 20	Method
[64]	2019	TrustBus *(conference)*	12	Domain understanding & elicitation			Architecture
[65]	2018	ESPRE *(workshop)*	11	Domain understanding & elicitation	Lawfulness, Fairness, and Transparency; Purpose Limitation; Data Minimization; Accuracy; Storage Limitation; Integrity and Confidentiality; Accountability	5	Tool
[66]	2019	QUATIC *(conference)*	9	Domain understanding & elicitation			Template
[16]	2018	OTM *(conference)*	9	Domain understanding & elicitation	Lawfulness, Fairness, and Transparency; Integrity and Confidentiality; Purpose Limitation; Data Minimization; Right to Rectification; Right to Erasure; Right to Object; Right to Restriction; Right to Data Portability	5, 17, 18, 19, 20, 21	Approach
[67]	2019	RE *(conference)*	9	Specification & documentation			Architecture
[34]	2019	SAC *(symposium)*	8	Domain understanding & elicitation	Consent	7	Approach
[68]	2018	SBES *(symposium)*	8	Specification & documentation	Privacy		Framework
[47]	2018	ISoLA *(symposium)*	4	Evaluation & negotiation	Privacy		
[69]	2018	MoDRE *(workshop)*	4	Specification & documentation			Framework
[70]	2020	Software and Systems Modeling (*Journal*)	3	Specification & documentation	Privacy; Integrity and Confidentiality	5	Method & Tool
[71]	2018	ICEIS *(conference)*	1	Specification & documentation			Methodology
[53]	2020	ICISSP *(conference)*	0	Domain understanding & elicitation	Integrity and confidentiality	5	Method
[52]	2019	REW *(workshop)*	0	Domain understanding & elicitation	Anonymization		Approach
[72]	2019	EuroSPI *(conference)*	0	Domain understanding & elicitation	Right of Access; Right to Rectification; Right to Erasure; Right to Object; Right to Restriction; Right to Withdraw Consent; Personal Data Breach Notification; Right to Data Portability; DPIA; Consent; Processor Management; Record of Processing Management; Automated Decision-Making; Data Protection by Design; Personal Data Management	7, 15, 17, 18, 19, 20	Model
[61]	2018	PoEM *(conference)*	0	Evaluation & negotiation	Privacy		Modeling language

processing of textual data forms [14], matching related requirements to avoid duplication and tracking requirements [53]. NLP can assist industry professionals with these manual and potentially repetitive tasks. For example, the NLP-based tool proposed by Kenney and Cooper [54] targets duplicate requirements by addressing their lexical and semantic similarity using NLP-techniques.

This mapping study shows that significant research has been conducted in RE and NLP; meanwhile, this trend is not (yet) apparent for the discussed paradigms related to the GDPR. However, as elaborated in the previous paragraphs, the results of this mapping study provide leads for narrowing the gap between GDPR, NLP, and RE, and thus achieving data protection by design through NLP-automated RE.

B. THREATS TO VALIDITY

This mapping study followed the guidelines as proposed by Keele [22] to identify all relevant research using a reliable, systematic, and rigorous methodology. However, since the mapping process is conducted manually, it is prone to human errors. In this Section, we will discuss the threats to the validity of this study.

1) STUDY SELECTION

It is possible that we disregarded relevant studies during the study selection process. The accuracy of the terms in our research queries impacts the relevancy of the results. We tried to overcome this by conducting several trial searches to perfect the selection of keywords. Furthermore, we carefully selected—through trial searches—different electronic databases endeavoring to increase the chances of gathering the most accurate set of studies. Finally, we approached the studies with a philosophy of caution by using a multi-step selection procedure. In the first step, the studies were excluded—if there was sufficient evidence—based on title, abstract, and introduction. However, if the studies did not indicate irrelevancy, they were reviewed in full.

2) MISCLASSIFICATION

Another threat to validity is the possibility of misclassification by the authors. Predilection, misreading, or misinterpreting an article may lead to a wrong classification. This is particularly the case for the data items that were classified subjectively, and not by explicit mentioning. This threat was mitigated by the scrupulous drafting of a mapping protocol and predefining—except for the NLP approaches—classifications of relevant data; and by preparing and validating a data extraction form through which the relevant data was collected. These measures were taken to reduce subjectivity in the mapping process.

VI. CONCLUSION

This paper has presented a systematic mapping study on GDPR, NLP, and RE to explore the literature on automated (i.e., using NLP) GDPR-compliance in RE. The combination of these three domains did not yield results according to the preliminary review. Therefore, three different research questions were designed to map the literature in the different pairwise combinations. In total, we have identified 448 studies conducted between 2010-2020. The majority, 420 studies, were collected to answer our first research question and identify NLP approaches employed in different RE activities. Our second research question focused on identifying NLP approaches used for GDPR-compliance in organizations and resulted in nine studies. Lastly, 20 studies were retrieved to address our third research question, which centers around RE solutions for achieving GDPR-compliance. Although only one of the retrieved studies was marked as combining GDPR, NLP, and RE, the mapping results indicate bridging opportunities between these three fields. In particular, we identified the possibilities of introducing NLP for automating manual tasks in the crossing of GDPR and RE, in addition to possibilities of using NLP-based machine learning techniques to achieve GDPR-compliance in RE. In combination with the indications for bridging the disciplines of GDPR, NLP, and RE, the findings of this systematic mapping study can be used as a stepping stone for both academia and industry. A comprehensive and clear mapping of all the retrieved studies can be found at https://aberkane.github.io/SMS_GDPR-NLP-RE.

APPENDIX A
DATA EXTRACTION FORM

DATA EXTRACTION FORM

See Table 9.

APPENDIX B
SELECTED STUDIES

LIST OF SELECTED STUDIES: RQ2

See Table 10.

LIST OF SELECTED STUDIES: RQ 3

See Table 11.

REFERENCES

[1] The European Parliament and the Council of European Union, "REGULATION (EU) 2016/679," *Off. J. Eur. Union*, pp. 1–2, Apr. 2016.

[2] B. Nuseibeh and S. Easterbrook, "Requirements engineering: A roadmap," in *Proc. Conf. Future Softw. Eng.*, May 2000, pp. 35–46.

[3] C. Tikkinen-Piri, A. Rohunen, and J. Markkula, "EU general data protection regulation: Changes and implications for personal data collecting companies," *Comput. Law Secur. Rev.*, vol. 34, no. 1, pp. 134–153, Feb. 2018.

[4] F. Dalpiaz, A. Ferrari, X. Franch, and C. Palomares, "Natural language processing for requirements engineering: The best is yet to come," *IEEE Softw.*, vol. 35, no. 5, pp. 115–119, Sep. 2018.

[5] M. Kassab, C. Neill, and P. Laplante, "State of practice in requirements engineering: Contemporary data," *Innov. Syst. Softw. Eng.*, vol. 10, no. 4, pp. 235–241, Apr. 2014.

[6] H. Li, L. Yu, and W. He, "The impact of GDPR on global technology development," *J. Global Inf. Technol. Manage.*, vol. 22, no. 1, pp. 1–6, Jan. 2019.

[7] European Commission. (2018). *The GDPR: New Opportunities, new Obligations Needs to Know About the EU's General Data Protection Regulation business*. Accessed: Jan. 1, 2021. [Online]. Available: https://ec.europa.eu/commission/sites/beta-political/files/data-protect%ion-factsheet-sme-obligations_en.pdf

[8] J. P. Albrecht, "How the GDPR will change the world," *Eur. Data Protection Law Rev.*, vol. 2, no. 3, pp. 287–289, 2016.

[9] P. Voigt and A. von dem Bussche, *The EU General Data Protection Regulation*. vol. 111, no. 5. Cham, Switzerland: Springer, 2017, p. 62.

[10] A. Van Lamsweerde, *Requirements Engineering: From System Goals to UML Models to Software*. vol. 10. Chichester, U.K.: Wiley, 2009, pp. 30–34.

[11] J. McManus and T. Wood-Harper, "A study in project failure," *Brit. Comput. Soc. Chartered Inst. IT*, pp. 1–4, Jun. 2008.

[12] E. D. Liddy, "Natural language processing," in *Encyclopedia of Library and Information Science*, 2nd ed. New York, NY, USA:Marcel Decker, 2001.

[13] D. Torre, G. Soltana, M. Sabetzadeh, L. C. Briand, Y. Auffinger, and P. Goes, "Using models to enable compliance checking against the GDPR: An experience report," in *Proc. ACM/IEEE 22nd Int. Conf. Model Driven Eng. Lang. Syst. (MODELS)*, Sep. 2019, pp. 1–11.

[14] V. Ayala-Rivera and L. Pasquale, "The grace period has ended: An approach to operationalize GDPR requirements," in *Proc. IEEE 26th Int. Requirements Eng. Conf. (RE)*, Aug. 2018, pp. 136–146.

[15] M. Robol, M. Salnitri, and P. Giorgini, "Toward GDPR-compliant socio-technical systems: Modeling language and reasoning framework," in *Proc. IFIP Work. Conf. Pract. Enterprise Model.*, Oct. 2017, pp. 236–250.

[16] S. D. Ringmann, H. Langweg, and M. Waldvogel, "Requirements for legally compliant software based on the gdpr," in *Proc. OTM Confederated Int. Conf. Move Meaningful Internet Syst.*, Oct. 2018, pp. 258–276.

[17] H. Meth, M. Brhel, and A. Maedche, "The state of the art in automated requirements elicitation," *Inf. Softw. Technol.*, vol. 55, no. 10, pp. 1695–1709, Oct. 2013.

[18] M. Bano, "Addressing the challenges of requirements ambiguity: A review of empirical literature," in *Proc. 5th Int. Workshop Empirical Requirements Eng.*, Aug. 2015, pp. 21–24.

[19] F. Nazir, W. H. Butt, M. W. Anwar, and M. A. K. Khattak, "The applications of natural language processing (NLP) for software requirement engineering—A systematic literature review," in *Proc. Int. Conf. Inf. Sci. Appl.*, Mar. 2017, pp. 485–493.

[20] M. Binkhonain and L. Zhao, "A review of machine learning algorithms for identification and classification of non-functional requirements," *Expert Syst. Appl., X*, vol. 1, Apr. 2019, Art. no. 100001.

[21] L. Zhao, W. Alhoshan, A. Ferrari, K. J. Letsholo, M. A. Ajagbe, E.-V. Chioasca, and R. T. Batista-Navarro, "Natural language processing (NLP) for requirements engineering: A systematic mapping study," 2020, *arXiv:2004.01099*. [Online]. Available: https://arxiv.org/abs/2004.01099

[22] S. Keele, "Guidelines for performing systematic literature reviews in software engineering," Dept. Comput. Sci., Univ. Durham, Durham, U.K., Tech. Rep. 2.3, Jul. 2007.

[23] D. Budgen, M. Turner, P. Brereton, and B. Kitchenham, "Using mapping studies in software engineering," in *Proc. 20th Annu. Meeting Psychol. Programm. Interest Group (PPIG)*, vol. 8, Jan. 2008, pp. 195–204.

[24] K. Petersen, R. Feldt, S. Mujtaba, and M. Mattsson, "Systematic mapping studies in software engineering," in *Proc. 2nd Int. Conf. Eval. Assessment Softw. Eng. (EASE)*, Jun. 2008, pp. 1–10.

[25] P. Brereton, B. A. Kitchenham, D. Budgen, M. Turner, and M. Khalil, "Lessons from applying the systematic literature review process within the software engineering domain," *J. Syst. Softw.*, vol. 80, no. 4, pp. 571–583, Apr. 2007.

[26] F. Dalpiaz and S. Brinkkemper, "Agile requirements engineering with user stories," in *Proc. IEEE 26th Int. Requirements Eng. Conf. (RE)*, Aug. 2018, pp. 506–507.

[27] H. Yin, "A study plan: Open innovation based on Internet data mining in software engineering," in *Proc. Int. Conf. Softw. Syst. Process*, Aug. 2015, pp. 192–193.

[28] J. Recker, *Scientific Research in Information Systems*. Berlin, Germany: Springer, 2013, pp. 36 and 52–56.

[29] C. Arora, M. Sabetzadeh, L. Briand, and F. Zimmer, "Automated extraction and clustering of requirements glossary terms," *IEEE Trans. Softw. Eng.*, vol. 43, no. 10, pp. 918–945, Oct. 2017.

[30] R. Wieringa, N. Maiden, N. Mead, and C. Rolland, "Requirements engineering paper classification and evaluation criteria: A proposal and a discussion," *Requirements Eng.*, vol. 11, no. 1, pp. 102–107, Mar. 2006.

[31] J. McZara, S. Sarkani, T. Holzer, and T. Eveleigh, "Software requirements prioritization and selection using linguistic tools and constraint solvers—A controlled experiment," *Empirical Softw. Eng.*, vol. 20, no. 6, pp. 1721–1761, Dec. 2015.

[32] S. Gregor and A. R. Hevner, "Positioning and presenting design science research for maximum impact," *MIS Quart.*, vol. 37, no. 2, pp. 337–355, Feb. 2013.

[33] M. Bano, A. Ferrari, D. Zowghi, V. Gervasi, and S. Gnesi, "Automated service selection using natural language processing," *Commun. Comput. Inf. Sci.*, vol. 558, pp. 3–17, Oct. 2015.

[34] C. Stach and F. Steimle, "Recommender-based privacy requirements elicitation—EPICUREAN," in *Proc. ACM Symp. Appl. Comput.*, Apr. 2019, pp. 1500–1507.

[35] European Commission. (2021). *Data protection in the EU*. Accessed: Feb. 7, 2021. [Online]. Available: https://ec.europa.eu/info/law/law-topic/data-protection/data-protection% -eu_en

[36] E. Guzman and W. Maalej, "How do users like this feature? A fine grained sentiment analysis of app reviews," in *Proc. IEEE 22nd Int. Requirements Eng. Conf. (RE)*, Aug. 2014, pp. 153–162.

[37] W. Maalej and H. Nabil, "Bug report, feature request, or simply praise? On automatically classifying app reviews," in *Proc. IEEE 23rd Int. Requirements Eng. Conf. (RE)*, Aug. 2015, pp. 116–125.

[38] W. Maalej, Z. Kurtanović, H. Nabil, and C. Stanik, "On the automatic classification of app reviews," *Requirements Eng.*, vol. 21, no. 3, pp. 311–331, Sep. 2016.

[39] G. Lucassen, F. Dalpiaz, J. M. E. M. van der Werf, and S. Brinkkemper, "Improving agile requirements: The quality user story framework and tool," *Requirements Eng.*, vol. 21, no. 3, pp. 383–403, Sep. 2016.

[40] H. Yang, A. de Roeck, V. Gervasi, A. Willis, and B. Nuseibeh, "Analysing anaphoric ambiguity in natural language requirements," *Requirements Eng.*, vol. 16, no. 3, pp. 163–169, Sep. 2011.

[41] N. Hariri, C. Castro-Herrera, M. Mirakhorli, J. Cleland-Huang, and B. Mobasher, "Supporting domain analysis through mining and recommending features from online product listings," *IEEE Trans. Softw. Eng.*, vol. 39, no. 12, pp. 1736–1752, Dec. 2013.

[42] X. Xiao, A. Paradkar, S. Thummalapenta, and T. Xie, "Automated extraction of security policies from natural-language software documents," in *Proc. ACM SIGSOFT 20th Int. Symp. Found. Softw. Eng. (FSE)*, Nov. 2012, pp. 1–11.

[43] D. K. Deeptimahanti and R. Sanyal, "Semi-automatic generation of UML models from natural language requirements," in *Proc. 4th India Softw. Eng. Conf. (ISEC)*, Feb. 2011, pp. 165–174.

[44] C. Arora, M. Sabetzadeh, L. Briand, and F. Zimmer, "Automated checking of conformance to requirements templates using natural language processing," *IEEE Trans. Softw. Eng.*, vol. 41, no. 10, pp. 944–968, Oct. 2015.

[45] D. Falessi, G. Cantone, and G. Canfora, "Empirical principles and an industrial case study in retrieving equivalent requirements via natural language processing techniques," *IEEE Trans. Softw. Eng.*, vol. 39, no. 1, pp. 18–44, Jan. 2013.

[46] Y.-S. Martin and A. Kung, "Methods and tools for GDPR compliance through privacy and data protection engineering," in *Proc. IEEE Eur. Symp. Secur. Privacy Workshops (EuroS&PW)*, Apr. 2018, pp. 108–111.

[47] G. Schneider, "Is privacy by construction possible?" in *Proc. Int. Symp. Leveraging Appl. Formal Methods*, Nov. 2018, pp. 471–485.

[48] Information Commisioners Office. *Information Rights at the end of the Transition Period—Frequently Asked Questions*. Accessed: Jan. 27, 2021. [Online]. Available: https://ico.org.uk/for-organisations/dp-at-the-end-of-the-transition-pe% riod/transition-period-faqs/

[49] M. Dias, J. C. Ferreira, R. Maia, P. Santos, and R. Ribeiro, "Privacy in text documents," in *Proc. 33rd Int. Bus. Inf. Manage. Assoc. Conf.*, Apr. 2019, pp. 2551–2560.

[50] F. Di Cerbo and S. Trabelsi, "Towards personal data identification and anonymization using machine learning techniques," in *Proc. Commun. Comput. Inf. Sci.*, vol. 909, Aug. 2018, pp. 118–126.

[51] C. Chang, H. Li, Y. Zhang, S. Du, H. Cao, and H. Zhu, "Automated and personalized privacy policy extraction under GDPR consideration," in *Proc. Int. Conf. Wireless Algorithms, Syst., Appl.*, Jun. 2019, pp. 43–54.

[52] E. C. Groen and M. Ochs, "CrowdRE, user feedback and GDPR: Towards tackling GDPR implications with adequate technical and organizational measures in an effort-minimal way," in *Proc. IEEE 27th Int. Requirements Eng. Conf. Workshops (REW)*, Sep. 2019, pp. 180–185.

[53] S. Zinsmaier, H. Langweg, and M. Waldvogel, "A practical approach to stakeholder-driven determination of security requirements based on the GDPR and common criteria," in *Proc. 6th Int. Conf. Inf. Syst. Secur. Privacy*, Jan. 2020, pp. 473–480.

[54] O. Kenney and M. Cooper, "Automating requirement quality standards with QVscribe," in *Proc. CEUR Workshop*, vol. 2584, 2020.

[55] C. Song and V. Shmatikov, "Auditing data provenance in text-generation models," in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2019, pp. 196–206.

[56] M. Blohm, C. Dukino, M. Kintz, M. Kochanowski, F. Koetter, and T. Renner, "Towards a privacy compliant cloud architecture for natural language processing platforms," in *Proc. 21st Int. Conf. Enterprise Inf. Syst.*, vol. 1, Jan. 2019, pp. 442–449.

[57] S. P. Nayak and S. Pasumarthi, "Automatic detection and analysis of DPP entities in legal contract documents," in *Proc. 1st Int. Conf. Digit. Data Process. (DDP)*, Nov. 2019, pp. 70–75.

[58] D. N. Jaidan, M. Carrere, Z. Chemli, and R. Poisvert, "Data anonymization for privacy aware machine learning," in *Proc. Int. Conf. Mach. Learn., Optim., Data Sci.*, Jan. 2019, pp. 725–737.

[59] A. Alamäki, L. Aunimo, H. Ketamo, and L. Parvinen, "Interactive machine learning: Managing information richness in highly anonymized conversation data," in *Proc. Work. Conf. Virtual Enterprises*, vol. 568, Aug. 2019, pp. 173–184.

[60] P. Silva, C. Gonçalves, C. Godinho, N. Antunes, and M. Curado, "Using natural language processing to detect privacy violations in online contracts," in *Proc. ACM Symp. Appl. Comput.*, Mar. 2020, pp. 1305–1307.
[61] M. Robol, E. Paja, M. Salnitri, and P. Giorgini, "Modeling and reasoning about privacy-consent requirements," in *Proc. IFIP Work. Conf. Pract. Enterprise Model.*, vol. 335, Oct. 2018, pp. 238–254.
[62] M. Alshammari and A. Simpson, "Towards a principled approach for engineering privacy by design," in *Proc. 5th Annu. Privacy Forum*, Oct. 2017, pp. 161–177.
[63] R. Meis and M. Heisel, "Computer-aided identification and validation of intervenability requirements," *Information*, vol. 8, no. 1, p. 30, Mar. 2017.
[64] L. Piras, M. G. Al-Obeidallah, A. Praitano, A. Tsohou, H. Mouratidis, B. Gallego-Nicasio Crespo, J. B. Bernard, M. Fiorani, E. Magkos, A. C. Sanz, M. Pavlidis, R. D'Addario, and G. G. Zorzino, "DEFeND Architecture: A privacy by design platform for GDPR compliance," in *Proc. 16th Int. Conf. Trust Privacy Digit. Bus.*, Aug. 2019, pp. 78–93.
[65] J. Coles, S. Faily, and D. Ki-Aries, "Tool-supporting data protection impact assessments with CAIRIS," in *Proc. 2018 5th Int. Workshop Evolving Sec. Privacy Requirements Eng.*, Aug. 2018, pp. 21–27.
[66] C. Bartolini, S. Daoudagh, G. Lenzini, and E. Marchetti, "Gdpr-based user stories in the access control perspective," in *Proc. Commun. Comput. Inf. Sci.*, vol. 1010, Aug. 2019, pp. 3–17.
[67] K. Hjerppe, J. Ruohonen, and V. Leppanen, "The general data protection regulation: Requirements, architectures, and constraints," in *Proc. IEEE 27th Int. Requirements Eng. Conf. (RE)*, Sep. 2019, pp. 265–275.
[68] M. M. Peixoto and C. Silva, "Specifying privacy requirements with goal-oriented modeling languages," in *Proc. ACM Int. Conf.*, Sep. 2018, pp. 112–121.
[69] A. Rabinia and S. Ghanavati, "The FOL-based legal-GRL (FLG) framework: Towards an automated goal modeling approach for regulations," in *Proc. 8th Int. Model-Driven Requirements Eng. Workshop (MoDRE)*, Sep. 2018, pp. 58–67.
[70] M. Salnitri, K. Angelopoulos, M. Pavlidis, V. Diamantopoulou, H. Mouratidis, and P. Giorgini, "Modelling the interplay of security, privacy and trust in sociotechnical systems: A computer-aided design approach," *Softw. Syst. Model.*, vol. 19, no. 2, pp. 467–491, Mar. 2020.
[71] M. Fernandes, A. R. Silva, and A. Gonçalves, "Specification of personal data protection Requirements–Analysis of legal requirements from the GDPR regulation," in *Proc. 20th Int. Conf. Enterprise Inf. Syst.*, vol. 2, Jan. 2018, pp. 398–405.
[72] S. Cortina, P. Valoggia, B. Barafort, and A. Renault, "Designing a data protection process assessment model based on the GDPR," *Commun. Comput. Inf. Sci.*, vol. 1060, pp. 136–148, Aug. 2019.

ABDEL-JAOUAD ABERKANE received the B.Sc. degree in computing science from the University of Amsterdam, Amsterdam, The Netherlands, in 2016, and the M.Sc. degree in information science from Utrecht University, Utrecht, The Netherlands, in 2018. He is currently pursuing the Ph.D. degree in automated GDPR-compliance in requirements engineering with the Business Informatics Research Group, Ghent University, Ghent, Belgium. His research interests include natural language processing, requirements engineering, and the general data protection regulation.

GEERT POELS is currently a Senior Full Professor of business informatics with the Faculty of Economics and Business Administration, Ghent University, Ghent, Belgium, where he teaches intermediate and advanced courses on information systems, IT management, enterprise architecture, and service design. He also teaches in the master's degree in enterprise ICT architecture with the IC Institute Beerzel, Belgium. He supervises Ph.D. research on digital marketplaces, cybersecurity, and GDPR. His research relates to conceptual modeling (as research method) and enterprise modeling (as research domain), with a focus on business process architecture mapping, ArchiMate, value modeling, and NLP-based automated generation of conceptual models out of user requirements documents. As in academic service, he was a member of the development team of the COBIT 2019 framework for IT governance.

SEPPE VANDEN BROUCKE received the Ph.D. degree in applied economics with the KU Leuven, Belgium, in 2014. He is currently working as an Assistant Professor with the Department of Business Informatics, Ghent University, Ghent, Belgium. His work has been published in well-known international journals and presented at top conferences. He has also coauthored several books, including *Principles of Database Management* (Cambridge University Press), *Practical Web Scraping for Data Science* (Apress), and *Beginning Java Programming* (Wiley). His research interests include business data mining and analytics, machine learning, process management, and process mining.