Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

School of Computer Science and Statistics

ADAPT Research Centre
Knowledge and Data Engineering Group (KDEG)

Representing Activities associated with Processing of Personal Data and Consent using Semantic Web for GDPR Compliance

Harshvardhan J. Pandit

Supervised by Prof. Dave Lewis
Co-supervised by Prof. Declan O'Sullivan

2020

A thesis submitted to the
University of Dublin, Trinity College
in fulfilment of the requirements of the degree of
Doctor of Philosophy

This page intentionally left blank.

© Harshvardhan J. Pandit, 2020
https://harshp.com/research/phd-thesis

This work is licensed under a Creative Commons "Attribution 4.0 International" license.

This page intentionally left blank.

DECLARATION

I declare that this thesis has not been submitted as an exercise for a degree at this or any other university and it is entirely my own work.

I agree to deposit this thesis in the University's open access institutional repository or allow the Library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement.

Signed.
Harshvardhan J. Pandit

This page intentionally left blank.

PERMISSION TO LEND OR COPY

I, the undersigned, agree that the Trinity College Library may lend or copy this thesis upon request.

Date:

Signed.
Harshvardhan J. Pandit

This page intentionally left blank.

ACKNOWLEDGEMENTS

The first and foremost acknowledgement I would like to make is in thanking my supervisors - *Dave Lewis* and *Declan O'Sullivan*. This work has been made possible by their vision, encouragement, motivation, and guidance.

I would also like to express my thanks to my parents - *pappa* and *mummy* - who always have and continue to provide support and motivation since I came to being. Thank you *Annalina Caputo* for your understanding, strength, and friendship in this journey. Thank you *Gauri Noolkar* for asking me to always keep moving ahead and for pushing me through the difficult times.

The environment at work and away was a joy thanks to the many people whom I met and were around. It is a pleasure to know you - *Gary, Nicole, Lucy, Ademar, Brendan, Jamie, Javier, Kris, Christoph, Anirban, Ramisa, Anuj, Dominika, Sahil, Judie, Jeremy*, and many others.

Thank you to *Sabrina, Axel, Javier* for giving me the opportunity to visit Vienna and work there. Thanks to *Rob, Christoph*, and *Eoin* for providing guidance and assistance, and to *Mark* for pushing me to apply myself in the real world.

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

This page intentionally left blank.

ABSTRACT

The General Data Protection Regulation (GDPR) is a landmark law regarding privacy and data protection. GDPR stipulates potentially large fines if an organisation is found to not be compliant. This has resulted in research involving use of technological resources to meet and evaluate compliance requirements. Such approaches involve representing information about processing activities in a machine-readable format, verifying its correctness, and evaluating whether it meets the obligations of GDPR.

Under GDPR, an organisation is required to maintain and demonstrate documentation showing its compliance to the obligations and requirements regarding processing activities. This documentation involves information on how the activities were planned, evaluated for compliance, and executed. In addition to these, if consent is used as a legal basis to justify the processing, then information about how that consent was obtained also needs to be recorded in order to evaluate and demonstrate its adherence to requirements specified by GDPR.

Utilising semantic web technologies provides a machine-readable and interoperable representation of information that can be queried and verified based on open standards such as RDF, OWL, SPARQL, and SHACL. This thesis presents the use of semantic web technologies to represent and associate information regarding processing of personal data and consent with GDPR for assistance with its compliance. In particular, it addresses three deficits within the current state of the art for utilising linked data approaches for GDPR compliance. The first of these is regarding associating information with the text and concepts of GDPR which would enable the adoption of a linked data approach to automation and management of compliance documentation. The second concerns representations of activities regarding the planning and execution of processes concerning personal data and consent. The third involves representing information required to evaluate and demonstrate compliance with the requirements of consent.

The outcomes of the work are presented in the thesis in the form of major contributions of GDPRtEXT - a linked data representation of the text of GDPR and a glossary of concepts relevant for its compliance, GDPRov - an OWL2 ontology based on PROV-O for modelling activities associated with personal data and consent in ex-ante (planning) and ex-post (execution) phases, and GConsent - an OWL2 ontology for representing information regarding consent. The thesis also presents minor contributions describing application of semantic web technologies in the form of querying and validation of information using the SPARQL and SHACL standards.

This page intentionally left blank.

CONTENTS

1 Introduction **1**

1.1 Background & Motivation . . . 1
1.2 Research Question . . . 4
1.2.1 Definitions . . . 5
1.2.2 Research Objectives . . . 5
1.3 Research Methodology . . . 7
1.3.1 Reviewing the State of the Art . . . 7
1.3.2 Information Gathering . . . 8
1.3.3 Ontology Engineering . . . 8
1.3.4 Querying Information for GDPR Compliance . . . 9
1.3.5 Information Validation Framework for GDPR Compliance . . . 9
1.3.6 Evaluation Methodology . . . 9
1.4 Contributions of this Thesis . . . 11
1.4.1 GDPR as a Linked Data Resource . . . 11
1.4.2 Ontologies for representing activities about Personal Data and Consent 12
1.4.3 Querying Information Related to Compliance using SPARQL . . . 13
1.4.4 Framework for Validating Information using SHACL Compliance . . . 14
1.4.5 Information Interoperability Model of the GDPR . . . 14
1.4.6 Participation in DPVCG . . . 15
1.4.7 Publications . . . 15
1.5 Thesis Overview . . . 19

2 Background: GDPR and the Semantic Web **23**

2.1 General Data Protection Regulation (GDPR) . . . 23
2.1.1 Terminology . . . 23
2.1.2 Transparency and Requirements . . . 24
2.1.3 Data Protection Impact Assessment . . . 25
2.1.4 Sources of Additional Information . . . 25
2.2 Semantic Web Technologies . . . 27
2.2.1 RDF, RDFS, and OWL . . . 27
2.2.2 SPARQL Protocol and RDF Query Language (SPARQL) . . . 28
2.2.3 Shapes Constraint Language (SHACL) . . . 28
2.2.4 Standardised Ontologies . . . 29

3 State of the Art **31**
3.1 Methodology . . . 33
3.1.1 Identification and classification of approaches . . . 33
3.1.2 Criteria for Analysis . . . 33
3.2 Approaches for GDPR compliance utilising Semantic Web . . . 34
3.3 Other approaches addressing GDPR compliance . . . 48
3.4 Approaches involving Privacy Policies . . . 60
3.5 Approaches related to Consent . . . 62
3.6 Upcoming research projects addressing GDPR . . . 63
3.7 Analysis . . . 66
3.7.1 Overview of SotA . . . 66
3.7.2 Representation of GDPR . . . 69
3.7.3 Representation of activities . . . 69
3.7.4 Representation of Consent . . . 70
3.7.5 Querying of information associated with GDPR Compliance . . . 72
3.7.6 Evaluation of GDPR Compliance . . . 72
3.8 Gaps and Opportunities for Further Work . . . 74

4 Analysing GDPR Compliance Requirements **77**
4.1 Interoperability Model of Information based on GDPR . . . 78
4.2 Compliance Questions . . . 81
4.2.1 Methodology . . . 82
4.2.2 List of Questions . . . 83
4.2.3 Assumptions & Constraints . . . 90

5 Representing Information for GDPR Compliance using Ontologies **99**
5.1 Methodology for Ontology Engineering . . . 99
5.1.1 Utilisation of Existing Ontology Engineering Methodologies . . . 99
5.1.2 Ontology Quality . . . 100
5.1.3 Ontology Documentation . . . 101
5.1.4 Dissemination . . . 101
5.1.5 Evaluation . . . 102
5.2 GDPRtEXT - Linked Open Dataset of GDPR text & Glossary of Concepts . . . 104
5.2.1 Motivation . . . 104
5.2.2 Ontology Engineering and Creation of Resource . . . 105
5.2.3 Resource Description & Application . . . 108
5.2.4 Evaluation . . . 117
5.3 GDPRov - Ontology for GDPR activities associated with Personal Data and Consent . . . 120
5.3.1 Identification of requirements from competency questions . . . 120
5.3.2 Extending PROV-O and P-Plan . . . 124
5.3.3 Ontology Description & Application . . . 127
5.3.4 Evaluation . . . 131
5.4 GConsent - Ontology of Consent Information for GDPR Compliance . . . 140
5.4.1 Distinction with existing work in state of the art . . . 140

5.4.2 Relationship with GDPRov . . . 141
5.4.3 Requirements Gathering and Establishment of Competency Questions 141
5.4.4 Ontology Description & Application . . . 143
5.4.5 Evaluation . . . 146
5.5 Data Privacy Vocabulary (DPV) . . . 153
5.5.1 Relevance of DPV to this thesis . . . 153
5.5.2 Overview of DPV . . . 154
5.5.3 Comparing DPV with GDPRtEXT, GDPRov, and GConsent . . . 156
5.5.4 Comparing DPV with SotA . . . 158

6 Querying and Validating Information for GDPR Compliance 161
6.1 Querying Information using SPARQL . . . 161
6.1.1 SPARQL queries and ontological representation of information . . . 161
6.1.2 Methodology . . . 162
6.1.3 Demonstration using synthetic use-case . . . 169
6.1.4 Evaluation . . . 170
6.2 Validating Information using SHACL . . . 172
6.2.1 Validation Model . . . 173
6.2.2 Creation of SHACL constraints . . . 174
6.2.3 Utilising ex-ante test results for ex-post validations . . . 177
6.2.4 Proof-of-concept implementation . . . 178
6.2.5 Generating reports using SPARQL . . . 182
6.2.6 Evaluation . . . 184

7 Conclusion 189
7.1 Fulfilment of Research Objectives . . . 189
7.2 Extent of semantic web technologies in addressing RQ . . . 192
7.3 Contributions . . . 193
7.4 Opportunities for Further Work . . . 196
7.5 Final Remarks . . . 199

References 201

List of Figures

3.1 Overview of SPECIAL Architecture [85] . . . 35
3.2 Overview of SPECIAL Policy Log Vocabulary [51] . . . 36
3.3 Extension of ODRL for representing GDPR obligations [40] . . . 37
3.4 CitySPIN's extension of SPECIAL vocabularies [93] . . . 39
3.5 Data Categories represented in PrOnto [55] . . . 41
3.6 Information Model Ontology in BPR4GDPR [124] . . . 43
3.7 Compliance meta-model ontology in BPR4GDPR [125] . . . 43
3.8 Ontology of GDPR, PCI DSS, and CSA by Elluri et al. [127] . . . 44
3.9 Data Privacy Policy by Joshi and Banerjee [128] . . . 45
3.10 GDPR Data Provenance Model by Ujcich et al. [50] . . . 46
3.11 Ontology for GDPR and Information Security by Geko & Tjoa [133] . . . 48
3.12 GuideMe approach by Ayala-Rivera and Pasquale [134] . . . 49
3.13 Consent and Data Management Model by Peras [137] . . . 50
3.14 Conceptual Model of GDPR by Tom et al. [138] . . . 51
3.15 PE-BPMN approach by Pullonen et al. [139] . . . 52
3.16 Architecture of LUCE [142] . . . 53
3.17 Meta-model of architectural view for GDPR by Sion et al. [143] . . . 55
3.18 meta-model for Privacy Level Agreements by Diamantopoulou et al. [146] . . . 57
3.19 Model of policy specification framework in RestAssured [153] . . . 59
3.20 Privacy taxonomy by Wilson et al. [167] used by Harkous et al. [164] and Linden et al. [165] to analyse privacy policies . . . 61
3.21 Informed consent permissions ontology by Grando et al. [169] . . . 62
3.22 Fields in Consent Receipt [170] . . . 63

4.1 Interactions between entities based on requirements of GDPR [67] . . . 79

5.1 Article 12(3) in GDPRtEXT as RDF displayed using Pubby [72] . . . 108
5.2 Visual overview of concepts in GDPRtEXT - part (a) [72] . . . 109
5.3 Visual overview of concepts in GDPRtEXT - part (b) [72] . . . 110
5.4 Overview of PROV-O model [47] . . . 125
5.5 Overview of P-Plan model and its relationship with PROV-O [48] . . . 126
5.6 GDPRov concepts derived by extending PROV-O and P-Plan . . . 128
5.7 Example steps depicting data life-cycle using GDPRov . . . 129
5.8 Consent life-cycle defined using GDPRov . . . 130
5.9 Modelling changes in workflows using P-Plan [46] . . . 132

5.10 Core concepts in GConsent [71] . . . 143
5.11 Concepts for representing context of consent in GConsent [71] . . . 144
5.12 Concepts representing state/status of consent in GConsent [71] . . . 145
5.13 GConsent representation of use-case involving third party data sharing [71] . 146
5.14 Core concepts in DPV [78] . . . 154

6.1 Questions for information required to assess compliance - Page 10 of "Preparing Your Organisation for the GDPR - A Guide for SMEs" published by Ireland's Data Protection Commission . . . 163
6.2 Retrieving information using SPARQL for query G5 in GDPR readiness checklist . . . 171
6.3 Overview of approach utilising SHACL to validate information for GDPR compliance . . . 173
6.4 Consent dialogues on quantcast.com (clockwise from top-left) (a) first screen (b) default options on selecting "I Accept" (c) default options on selecting "Show Purposes" (d) Third parties listed for purpose "Personalisation" . . . 179

List of Tables

1.1 Summary of Evaluation Methods 9

3.1 Overview of approaches in SotA 67
3.2 Approaches representing GDPR in machine-readable format 69
3.4 Representation of activities in SotA 70
3.6 Representation of consent in SotA 71
3.8 Compliance evaluation in SotA 73

5.1 Concepts in GDPRtEXT for answering competency questions 117
5.2 Comparison of GDPRtEXT with SotA 119
5.3 Concepts in GDPRov for answering competency questions 133
5.4 Comparison of GDPRov with SotA 135
5.5 GDPRov concepts to represent external use-case from SPECIAL 136
5.6 Concepts in GConsent for answering competency questions 147
5.7 Comparison of GConsent with SotA 149
5.8 GConsent concepts to represent external use-case from SPECIAL 150

6.1 Questions provided in the GDPR Readiness Guide 164
6.2 Analysis of compliance questions specified in Table 6.1 166
6.3 SHACL validation report linked to GDPR 183
6.4 Comparison of SHACL validation with SotA 186

LIST OF SOURCE CODES

1 Use of GDPRtEXT to link tests with GDPR Articles in EARL report 115
2 SPARQL query and results showing retrieved GDPR test results by article . . 115
3 Example annotation of associating existing DPD compliance XACML rules with requirements of GDPR 116
4 SPARQL query representing compliance question `G5` concerning legal basis for processing 131
5 GDPRov representation of external use-case from SPECIAL 137
6 SPARQL queries using GDPRov for external use-case from SPECIAL 138
7 GConsent representation of external use-case from SPECIAL 151
8 SPARQL queries using GConsent for external use-case from SPECIAL 152
9 SPARQL query representing compliance question `G5` concerning legal basis for processing 169
10 Extending SHACL `NodeShape` to express manual and automated checking of constraints 175
11 SHACL constraint checking Data Subject associated with consent 176
12 Expressing the same constraint in SHACL-SPARQL and in SHACL core . . . 176
13 Evaluating manually checked constraints using boolean values 177
14 Utilising ex-ante test results for consent model in evaluating ex-post instances of given consent 178
15 Representation of consent dialogue as a bundle of consent requests 180
16 SPARQL query for report listing validation results linked with GDPR 183

List of Abbreviations & Acronyms

CCPA	California Consumer Protection Act
DPD	Data Protection Directive
DPV	Data Privacy Vocabulary
DPVCG	W3C Data Protection Vocabularies and Controls Community Group
ELI	European Legislative Identifier
EU	European Union
GDPR	General Data Protection Regulation
P-Plan	P-Plan Ontology
PROV-O	Provenance Ontology
ODRL	Open Digital Rights Language
OWL	Web Ontology Language
RDF	Resource Description Framework
SHACL	Shapes Constraint Language
SPARQL	SPARQL Protocol and RDF Query Language
SotA	State of the Art

List of RDF Namespace Prefixes

gc	https://w3id.org/GConsent#
gdprov	https://w3id.org/GDPRov#
gdprtext	https://w3id.org/GDPRtEXT#
pplan	http://purl.org/net/p-plan#
prov	http://www.w3.org/ns/prov#
rdf	http://www.w3.org/1999/02/22-rdf-syntax-ns#
rdfs	http://www.w3.org/2000/01/rdf-schema#

1 | INTRODUCTION

1.1 BACKGROUND & MOTIVATION

To date, 132 of the 206 states listed by the United Nations (UN) have a privacy law which regulates the usage of personal data [1]. However, their intended application suffers from a disconnect with the rapid progress in technology. In particular, the use of internet as a medium for data exchange and its pervasiveness and connectivity to individuals via devices such as the smartphone has led to industrial data harvesting at large scales [2]. To counter this problem, lawmakers in the European Union (EU) passed the General Data Protection Regulation (GDPR) [3] in 2016 with the aim of providing individuals with the right to information and control over use of their personal data, and to simplify requirements for organisations through a unified regulation across the EU.

The GDPR has received a large amount of attention due to its prospective fines which can potentially be up to 4% of an organisation's annual turnover or €20 million - whichever is greater. As of February 2020, there have been over 215 publicly known instances of fines associated with the GDPR [4], the largest of which was the €50 million fine to internet giant Google [5]. Being a regulation and replacing the Data Protection Directive (DPD) [6], GDPR provides a uniform set of compliance requirements across the EU, and is the basis of national privacy laws implemented in its member states [7]. Furthermore, GDPR has influenced other privacy laws, such as the California Consumer Protection Act (CCPA) [8], thereby further expanding similarities in compliance requirements across the globe.

The most visible change of the GDPR for most individuals is the ubiquitous 'consent dialogue' on websites that requests 'consent' - one of the legal basis for processing of personal data in GDPR. Despite being a legal requirement, consent dialogues have been accused of being non-transparent and subverting the spirit of the GDPR [9], [10]. The issue of consent itself has received significant interest in development and utilisation of technological solutions for compliance due to the right to withdraw consent provided by the GDPR which enables an individual to revoke their previously given consent and requires processing of personal data based on it to be halted. Opinions published by legal experts and bodies, in absence of case law on this issue, have expressed the need for greater transparency regarding activities associated with use of consent [11].

Compared to other privacy laws, including its predecessor DPD, GDPR provides significantly stricter and detailed requirements for processing of personal data and requires organisations to explicitly document information in relation to its obligations in order to be compliant. This information consists of identification of GDPR clauses applicable to the practices of an organisation and the steps taken to fulfil requirements and obligations for

compliance. From a technical or information management point of view, GDPR specifies interactions between entities in a clear manner. An example of this is an organisation using consent as the legal basis being required to provide information about processing activities to the data subject. Furthermore, this information is required to be maintained, evaluated, and documented to demonstrate compliance upon request by authorities. At the same time, this information is also associated with other stakeholders - such as through privacy policies, user agreements, terms and conditions, or even data processing agreements. This makes it clear that information associated with GDPR compliance is also used in other applications and involves multiple stakeholders.

As GDPR is a data protection law, its compliance is concerned primarily with processing of personal data, its legality, and associated operations within an organisation. This includes processing in both tenses - past as well as future - where an organisation is obligated to first determine and ensure its requirements and activities involving processing of personal data are valid as per the GDPR, and to then maintain a record of such activities as the processing takes place. These are defined[1] within the legal domain by the terms 'ex-ante' to specify compliance assessment before activity takes place (preventative) and 'ex-post' to specify compliance assessment after the activity has taken place (corroborative).

While the GDPR does not explicitly mention a 'phase' of compliance, its use enables associating the information to the planning and processing operations carried out within an organisation. The planning of processing operations also involves investigation of whether the intended operations will be compliant to the legal obligations, and the required corrections to ensure they continue to be so. The processing operations carried out also need to be inspected to ensure they met the requirements set forth in the planning stage and that the processing itself was legally compliant.

The combination of new requirements and significant fines has provided an incentive to utilise technology in meeting the obligations and requirements stipulated by GDPR towards its compliance. Existing efforts, such as the International Organization for Standardization (ISO), have addressed this change by updating standards to meet increased requirements with global privacy laws. In the context of GDPR, ISO/IEC 27001[2] defines requirements for an information security management system, and its extension ISO/IEC 27701[3] defines a privacy information management system, which together provide a framework for managing privacy risks associated with personal data processing. Adherence to such standards provides a commonality in the information management practices of an organisation, and assists in the compliance process by providing a structured interpretation and demonstration of practices based on the standardised specifications.

Technological development of solutions for legal compliance face two problems in general - the first being algorithmic interpretation of requirements associated with legal compliance. This is difficult as the text used in a legal document such as GDPR does not readily lead to algorithmic compliance due to ambiguity and uncertainty in its legal interpretation - especially in domain specific use-cases. In addition, because GDPR has been enforced for a comparatively short period - the interpretation of its clauses as requirements for compliance relies on clarification through legal opinions and decisions by supervisory authorities and

[1] Defined in EU terminology database (IATE) https://iate.europa.eu/entry/result/787324/en
[2] https://www.iso.org/standard/54534.html
[3] https://www.iso.org/standard/71670.html

courts. The second problem is that regardless of how technology is used in the compliance process, formal investigations of legal compliance require information to be documented and associated with the specifics of the law they intend to comply with - in this case the articles and clauses of GDPR. Traditionally, this is carried out through creation of documentation by legal experts, lawyers, and legal departments. Therefore, technological solutions addressing GDPR compliance must also provide information documentation in addition to assessment of compliance.

Incorporating legal compliance into organisational requirements has led to several approaches such as: use of symbolic (mathematical) logic, knowledge representation of legal text as logical rules, deontic rights specifying rights and obligations, defeasible logic based on exceptions, first order temporal logic, access control, markup based representations, and goal modelling of obligations [12]. While there has been significant work in the use of technology to adopt these approaches towards addressing and evaluating compliance in the last decade [12]–[18], the issue of associating information with legal documents has received relatively less attention. Where contemporary methods are sufficient to meet legal requirements, their use of text-based document formats prevents effective technological solutions that can be scaled, automated, or utilised in an information management system. To enable such approaches, information associated with compliance must be represented using machine-readable formats that enable the use of querying to retrieve information as well as validation methods to ensure its correctness. Furthermore, the need to share information between stakeholders defined within the GDPR provides motivation towards developing interoperability in information and solutions - which also provides transparency in the compliance process. By using open and interoperable standards, the commonality in representation and interpretation of information benefits stakeholders and reduces costs associated with innovation regarding information management and regulatory compliance.

Governmental agencies across the globe have addressed the issue of information interoperability by adopting the principles of Linked Data [19] and have produced interoperable standards [20]–[22] to facilitate use of information in technological solutions. These standards implement principles of the Semantic Web [23] by utilising the Resource Description Framework (RDF) [24] to specify information in an interoperable, extensible, and machine-readable manner. This has paved the way for development of technologies that address challenges associated with legal compliance through greater use of automation and operations at large scales. Consequently, the use of Linked Data and Semantic Web within the legal domain has resulted in the development of ontologies for organising and structuring information, reasoning and problem solving, semantic indexing and search, semantic integration and interoperability, and understanding the domain [25].

Semantic Web is also being used to address the challenges associated with GDPR compliance through commercial[4] solutions as well as large-scale European research projects such as SPECIAL [26], MIREL [27], DAPRECO [28], BPR4GDPR [29], and RestAssured [30]. The technological solutions developed within these utilise ontologies to represent the information required for compliance and a corresponding approach that expresses and evaluates obligations to assess compliance. In general, semantic web technologies provide numerous

[4]Example: Top Quadrant's Semantic Data Governance for GDPR Compliance https://www.topquadrant.com/gdpr-compliance/

advantages, such as: their basis in standards maintained through community and stakeholder engagement by W3C; an interoperable information representation specification that serves as the base for other specifications that build on top of it to provide knowledge modeling, querying, and validation; and the ability to easily build upon an existing knowledge-system by extending the underlying data models while retaining compatibility. These advantages make semantic web technologies and ideal and useful toolset in the legal compliance domain, and especially the GDPR given the emphasis of its requirements on information and documentation for compliance.

The work presented in this thesis also utilises the Semantic Web to address GDPR compliance. It focuses on the representation of activities associated with processing of personal data and consent as a subset of information relevant to the investigation of GDPR compliance. These activities correspond to how organisations plan their processing of personal data and execute or implement them, and are therefore relevant to the planning and management of operations within organisations. This includes activities associated with acquiring consent owing to the role of consent as a legal basis and the assertion that consent itself is also personal data. The novelty of this work lies in the application of linked data principles to associate information with GDPR and the advantages this provides in utilising semantic web technologies to represent, query, and validate information relevant for compliance.

The role of semantic web in this is towards representing information relevant for GDPR compliance that can be associated with the text of GDPR following Linked Data Principles. This involves use of existing standards of RDF [24] and Web Ontology Language (OWL2) [31] to represent information as ontologies, SPARQL [32] for querying information, and Shapes Constraint Language (SHACL) [33] to validate information. The use of semantic web standards and technologies enables the information to be persisted in a machine-readable, interoperable, and queryable form - and thus readily lends itself to automation using technological solutions in the areas of legal compliance and its documentation.

In terms of scope, the work presented in this thesis addresses only the representation and management of information associated with GDPR compliance, and is not intended to provide an authoritative assessment of compliance as only supervisory authorities and courts have legal authority in this matter. In the same vein, the research presented in this thesis is also not intended to replace professional opinions such as that offered by lawyers and legal experts. Instead, the intention of the work is to demonstrate the applicability and feasibility of using technology as a tool to assist with the compliance process.

1.2 Research Question

The research question investigated in this thesis is:

Research Question

To what extent can information regarding activities associated with processing of personal data and consent be represented, queried, and validated using Semantic Web technologies for GDPR compliance?

1.2.1 Definitions

The following definitions are used in the context of the research question outlined above and this thesis:

- *information regarding activities*: information about how processes, services, tasks, or other similar concepts are planned, executed or carried out, along with the resulting outcomes and the artefacts used or required;
- *activities associated with processing of personal data*: information about how personal data will be or has been obtained (its source), its usage - including storage, sharing, analysis, or other forms of processing;
- *activities associated with consent*: information about how consent will be or has been obtained, its usage as a legal basis, the information represented by consent, and its planned or recorded withdrawal;
- *querying*: retrieving information using a structured representation based on the underlying representation of information;
- *validation*: assessment of information to meet a constraint or requirement;
- *associate or link information with GDPR*: to establish an association or link between information and clauses or concepts of the text of GDPR;
- *subset of GDPR*: a subset of the clauses defined in the text of the GDPR;
- *ex-ante compliance*: compliance regarding processing before it has taken place, i.e. *A-priori*;
- *ex-post compliance*: compliance regarding processing after it has taken place, i.e. *A-posteriori*;
- *compliance questions*: questions that retrieve information relevant for determination of compliance;
- *transparency of information*: specifying or providing information in a way that enables others (external entities) to understand and analyse it.

1.2.2 Research Objectives

The research question represents a broad investigation which is difficult to address as a whole. Therefore, it is reconstructed as multiple 'sub-research questions' which are smaller in scope and provide specific aims in the form of research objectives. These objectives are influenced by the analysis of the state of the art and subsequent identification of gaps in Section 3.7 as potential opportunities to answer the research question. The first two objectives are structured on the identification of information required for GDPR compliance. The third objective focuses on the use of semantic web technologies for information representation, while the fourth and fifth objectives are associated with querying and validation of information respectively.

The GDPR is a legal document structured into 173 Recitals, 99 Articles, and 21 Citations. Of these, not all clauses are relevant to activities associated with personal data and consent. Therefore, the first research sub-question concerns investigation and identification of the sub-set of GDPR regarding activities associated with personal data and consent, along with information on the ex-ante and ex-post aspects of such activities towards compliance. This provides the first objective as:

$RO1$: Identify the subset of GDPR relevant for activities associated with processing of personal data and consent regarding compliance.

Following identification of the relevant sub-set of GDPR, information required to represent activities needs to be identified through 'compliance questions' representing an investigation process to identify the actors, entities, and relationships relevant for GDPR compliance. This provides the second objective as:

$RO2$: Identify information required to represent activities associated with processing of personal data and consent in investigation of GDPR compliance.

The identified information is then represented as semantic web ontologies consisting of concepts and relationships. This representation acts as the information model upon which questions or queries can be executed to retrieve information for determining compliance. The formalisation of information as an ontology provides a controlled vocabulary for validation of information to determine its sufficiency and correctness before determining compliance.

Instead of representing all required information in a single large ontology, modular ontologies provide better reuse and are easier to engineer [34]. A modular ontology is limited in scope towards representing a specific information category, and therefore is more consistent in its representation of concepts, and is easier to evaluate as compared to a larger ontology in which different concepts may have differing semantic connotations. Modular ontologies also provide better motivation for reuse though selective choosing of concepts in a module without dependency of concepts in other modules.

With this as motivation, the larger objective of $RO3$ for creating an ontology to address the research question is divided into three modular ontologies of: $RO3(a)$ - associating information with clauses and concepts of GDPR; $RO3(b)$ - representing information about activities associated with processing of personal data and consent; and $RO3(c)$ representing information about consent. This provides the third objective as:

$RO3$: Create OWL2 ontologies for representation information about:
(a): concepts and text of GDPR
(b): activities associated with processing of personal data and consent
(c): consent required to determine compliance

'Compliance questions' retrieve information required to determine compliance, and are important in the documentation process. The information retrieval can be automated by utilising SPARQL queries to represent compliance questions using corresponding concepts and relationships from the developed ontologies. This provides the fourth objective as:

$RO4$: Represent compliance questions using SPARQL to query information about activities associated with processing of personal data and consent

The determination of compliance includes assessing whether a given information satisfies all obligations and requirements, and also involves validation of information itself in terms of correctness and completeness. In software engineering processes such assessments are automated as 'tests' that validate data and produce a report to record documentation. The same principle is utilised here to assess information for correctness and completeness based on requirements of GDPR. This is done using SHACL which enables expressing validation requirements over developed ontologies and produces a report which can be persisted and linked back to the GDPR for documentation of compliance. This provides the fifth objective as:

> $RO5$: Utilise SHACL to:
> **(a)**: validate information for GDPR compliance regarding activities associated with processing of personal data and consent
> **(b)**: link validation results with GDPR

1.3 Research Methodology

1.3.1 Reviewing the State of the Art

A review of the state of the art (SotA) regarding approaches towards GDPR compliance was conducted at several stages of the research from March 2016 to September 2019. Publications associated with research objectives were driving factors in providing requirements to conduct a SotA review to capture the approaches and progress at that particular time. In addition, a general review of legal models for compliance was also conducted to identify relevant approaches which could be reused towards addressing requirements of the GDPR. The inclusion of approaches in SotA largely focused on the use of semantic web technologies and the extent of their applicability towards addressing the requirements of the GDPR.

An understanding of GDPR was obtained from sources including the official text of GDPR [3], its interpretation and clarification provided by authoritative bodies such as Data Protection Commissions in various jurisdictions, Article 29 Working Party (A29WP), and the European Data Protection Board (EDPB). In addition, guides and expert opinions provided by legal experts and organisations were utilised as non-authoritative sources to better understand requirements of GDPR compliance. Information requirements associated with compliance presented within the thesis are based on these sources and through studying case law related to interpretation of the GDPR where accessible.

Approaches and resources within SotA were reviewed where information was open and accessible - such as through academic publications and project deliverables. Where such information was not accessible - such as in commercial products and some resources in academic projects - only the available information was included in the review of SotA. Publications and resources were discovered through Google Scholar, Scopus, IEEExplore, ACM Digital Library, and through events such as conferences and events, and through dissemination networks such as Twitter. Zotero was used as a bibliography tool for managing references and notes.

1.3.2 Information Gathering

The gathering of information regarding requirements of GDPR and its compliance was done through a literature review of official and authoritative documentation published by legal bodies and organisations. In order to understand the requirements of GDPR and stakeholders involved, a model was developed to understand requirements for information interoperability for each stakeholder. The information about GDPR and its requirements was used to create 'compliance questions' to guide the ontology development process by acting as 'competency questions' (see Section 1.3.3) and to act as queries for retrieving information relevant to the compliance process. The questions also provided the basis for creating information validation constraints. This process fulfilled research objectives $RO1$ and $RO2$ and is described in Chapter 4.

1.3.3 Ontology Engineering

The ontologies developed to fulfil research objective $RO3$ used methodologies commonly adopted and recommended within the semantic web community. A general introductory guide for creating ontologies [35] was used to understand and start the process of ontology engineering. The actual construction of ontologies followed a combination of NeON methodology [34] and UPON Lite [36] - where NeON was used to identify existing scenarios and gather requirements and UPON Lite was used to derive actionable steps or tasks to build and test the ontology using an agile development process. The combination provided a methodology for identifying relevant information from the GDPR (using NeOn) and iteratively building and updating an ontology to represent it (using UPON Lite). The methodology is described in more detail in Section 5.1, with a summary as:

1. Identification of aims, objectives, scope
2. Identify and analyse relevant information
3. Create use-cases and competency questions
4. Identify concepts and relationships
5. Create Ontology
6. Evaluate
7. Progressive iterations following steps 2 to 6
8. Dissemination

Each ontology was documented with metadata based on best practices advocated by the semantic web community[5] for automatic generation of documentation using the WIDOCO tool [37]. The namespace IRI was defined with persistent identifiers through the use of W3ID[6]. The ontology itself was archived in the public open repository Zenodo[7] which provided it with DOIs. All code and resources associated with the ontologies are published in GitHub - an open and public code repository. The ontology and related resources were hosted on Trinity College Dublin servers to enable resolution of their IRIs on the internet.

[5] https://dgarijo.github.io/Widoco/doc/bestPractices/index-en.html
[6] http://w3id.org/
[7] https://zenodo.org/

1.3.4 Querying Information for GDPR Compliance

The querying of information utilised SPARQL and fulfilled research objective $RO4$. The methodology to represent compliance questions as SPARQL queries utilised questions from a real-world document published by the Irish Data Protection Commission for assisting organisations in evaluation their readiness for GDPR. The querying was demonstrated by representing each question within the document as a SPARQL query using the developed ontologies and executed over a synthetic use-case.

1.3.5 Information Validation Framework for GDPR Compliance

In order to demonstrate the validation of information, a modular framework was proposed in Section 6.2 consisting of creating a 'compliance graph' separate from the data graph for storing information relevant to compliance. This facilitated the querying and validation of information associated with compliance in a modular approach using SPARQL and SHACL respectively. The constraints and assumptions created from constraint questions in Chapter 4 were represented using SHACL and used to validate information based on obligations and requirements of GDPR compliance. Its application was demonstrated through a use-case evaluating validity of consent in a real-world website.

1.3.6 Evaluation Methodology

A summary of evaluations methods used in the thesis is presented in Table 1.1

Table 1.1: Summary of Evaluation Methods

Method	GDPRtEXT Ontology	GDPRov Ontology	GConsent Ontology	Querying using SPARQL	Validation using SHACL
Fulfilment of Competency Questions	✓	✓	✓	N/A	N/A
Semantic reasoner logical consistency	✓	✓	✓	✓	✓
OOPS! common pitfalls detection	✓	✓	✓	N/A	N/A
Documentation metadata and quality	✓	✓	✓	N/A	N/A
Demonstrate application to use-case	✓	✓	✓	✓	✓
External use-case	✗	✓	✓	✓	✓
Comparison with SotA	✓	✓	✓	✓	✓
Analysis of citations	✓	✓	N/A	✓	N/A
Dissemination of work (for providing transparency)					
Peer-reviewed publication	✓	✓	✓	✓	✓
Reproducibility (open access resources)	✓	✓	✓	✓	✓

1.3.6.1 Evaluating Ontologies

The developed ontologies presented in Chapter 5 were assessed in their sufficiency and completeness to answer the competency questions they were designed for. In addition, use-cases related to situations differing in compliance requirements were used to assess the ontology in terms of sufficient representation of related information. These use-cases were compiled from GDPR-related case law, SotA, and synthetic situations, and validated regarding information requirements with a legal expert. Ontologies were also evaluated using best

practices advocated by the community throughout its development using a semantic reasoner to ensure logical consistency in expressed facts and axioms, and by using the OOPS! [38] online service to detect common pitfalls in ontology design. Finally, the sections in this thesis describing each ontology present a comparison against similar ontologies identified in SotA to analyse novelty, strengths, and weaknesses. The sections also present relevant peer-reviewed publications where ontologies were presented and discussed. Citations to these publications were used to identify relevant approaches and to investigate criticisms and comparisons with other ontological representations.

An ad-hoc evaluation of ontologies is also presented through their use in querying and validation of information for research objectives $RO4$ and $RO5$. This demonstrated the sufficiency of each ontology to provide sufficient concepts for representing information to facilitate querying and validation processes.

1.3.6.2 Evaluating Querying of Information

The use of SPARQL to query information based on compliance question as presented in Section 6.1 was evaluated by applying it to questions in a document published by the Irish Data Protection Commission. The SPARQL queries utilised developed ontologies to represent the given question as a compliance question and provided an opportunity to evaluate the extent to which the ontology could represent these concepts. The approach itself was evaluated based on the extent to which the questions in the document could be expressed as SPARQL queries. Where a question could not or was not expressed using SPARQL, an analysis was carried out to determine the reason - such as the question not being in scope of the research question.

1.3.6.3 Evaluating Information Validation Framework

The framework developed for validating information using constraints derived from compliance questions is presented in Section 6.2. Its evaluation consisted of generating a synthetic use-case using the consent mechanism of a real-world website where the constraints related to consent and personal data activities were validated on information from the website . The use-case enabled representation of activities in both ex-ante and ex-post phases where ex-ante represented validity of the consent dialogue being presented, and ex-post represented determining validity of given consent. The information regarding activities related to personal data and consent within the use-case was represented using developed ontologies. SHACL was then used to define constraints derived from competency questions with links to GDPR added to the constraints using developed ontologies and custom properties.

The evaluation consisted of demonstrating use of SHACL and developed ontologies to express the constraints and the ability to link constraints and its validation results with relevant clauses of GDPR. The approach also demonstrated the use of validation results as actionable tasks for compliance associated with clauses of GDPR. The framework and the application were compared with approaches within the SotA to demonstrate novelty in use of SPARQL and SHACL for GDPR compliance.

1.4 Contributions of this Thesis

The two major contributions of this thesis are (based on ontologies in $RO3$): first - enabling association of information with the text of GDPR following linked data principles, and second - ontologies for representing information about activities associated with processing of personal data and consent. Minor contributions include formulating an information model of entities and their relationships in GDPR (based on information in $RO1$ and $RO2$), using semantic web technologies for querying (based on $RO4$) and validating (based on and $RO5$) information required for compliance. Resources associated with the contributions[8], including published papers[9], have been made accessible under open licenses (MIT, CC-by-4.0) for reproducibility and to foster adoption and re-use by the community.

1.4.1 GDPR as a Linked Data Resource

The first major contribution of this thesis is the GDPRtEXT resource - which provides a linked data version of the text of GDPR and a glossary of its concepts. It fulfils research objective $RO3(a)$ and enables fulfilment of $RO5(b)$ by exposing each individual article or point within the text of GDPR as a unique resource using semantic web to enable links to be established between information and clauses of the GDPR. As these links are machine-readable, they can be used in approaches to automate the generation and querying of information associated with GDPR - such as for compliance, management of business processes, or generation of privacy policies. Furthermore, GDPRtEXT extends and is therefore compatible with the ELI ontology [39] used by the European Publications Office to publish legislations - including GDPR. ELI currently provides representations only at the document level, which GDPRtEXT extends for representing clauses at a granular level. GDPRtEXT thus provides its features in a manner compatible and interoperable with ELI.

It is currently common practice to refer to concepts within legal documents such as GDPR by associating them with their defining or relevant clauses within the document. GDPRtEXT provides a glossary (and vocabulary) of concepts defined or referred to within GDPR to assist with use of concepts associated with its compliance. Each concept or term is associated with its definition or articles of relevance within GDPR by using the linked data version of text provided by GDPRtEXT. This provides another way to link information to GDPR through use of concepts and has been used to indicate the source in definitions of terms and relationships within the other developed ontologies (see Section 1.4.2).

GDPRtEXT fills an important gap in the state of the art (as investigated in Chapter 3) by providing a mechanism to link information with the text of GDPR in a machine-readable manner. It is the only provider of a semantic web glossary of terms associated with GDPR and its compliance with a reference to their definition and usage within the text of GDPR. While there are other comparable and relevant methods to address such information [40], [41], GDPRtEXT is currently the only one that uses and extends ELI [42] - the official metadata standard for European legislation documents, and is also the only open and accessible ontology regarding GDPR and its concepts [43].

[8] http://openscience.adaptcentre.ie/res/
[9] https://openscience.adaptcentre.ie/publications/

GDPRtEXT has been released[10] under an open license (CC-by-4.0) and has been incorporated into Ireland's open data portal[11]. The provision of machine-readable concepts and reference to clauses of the GDPR makes GDPRtEXT an important resource for use in legal knowledge graphs.

1.4.2 Ontologies for representing activities about Personal Data and Consent

The second major contribution of this thesis are the two semantic web ontologies - GDPRov for representing information about activities associated with processing of personal data and consent, and GConsent for representing information associated with determining compliance of consent. Both ontologies define concepts and relationships using GDPRtEXT to indicate source within GDPR.

Together with GDPRtEXT, GDPRov and GConsent enable representation of activities required to evaluate and validate compliance with the GDPR. Apart from advancing state of the art, the ontologies also provide a vocabulary of terms and concepts relevant for GDPR compliance, and demonstrate the use of legal documents as a source for ontologies using linked data principles. Their usefulness has been demonstrated in approaches of: representation of information in privacy policies [44], generation of privacy policies from metadata [45], and automating change-detection and its effects on activities [46]. GDPRov[12] and GConsent[13] are published under an open license (CC-by-4.0).

1.4.2.1 GDPRov

GDPRov enables representation of the processes and activities associated with the life-cycle of personal data and consent, and fulfils the research objectives $RO3(b)$ and $RO3(c)$. GDPRov extends PROV-O [47] - which is the W3C standard for defining provenance information - to define ex-post (activity logs indicating things that have happened) information, and P-Plan [48] to define ex-ante (as an abstract model, template, or plan) representations of PROV activities based on scientific workflows. This enables it to represent planned activities as a model or template which is required to assess ex-ante compliance, and to associate it with its corresponding executions which are required to assess ex-post compliance. The linking of information between ex-ante and ex-post phase in GDPRov comes from its basis in scientific workflows. It also provides the opportunity to exploit this association for a more efficient approach in evaluation of compliance, as proposed and demonstrated in Section 6.2, and summarised as a contribution in the sections below.

The state of art contains ontologies for representing activities and their provenance related to the GDPR [41], [49], including those utilising PROV [50], [51], and holistic approaches combining ex-ante and ex-post compliance [52]. In comparison, GDPRov provides the most exhaustive vocabulary of concepts based on the GDPR (based on comparisons demonstrated in Chapter 5), and is the only ontology to provide ex-ante and ex-post concepts within the same ontology. GDPRov thus advances the state of the art by providing the most comprehensive vocabulary for modelling and representing activities based on GDPR

[10] https://w3id.org/GDPRtEXT
[11] https://data.gov.ie/dataset/gdprtext
[12] https://w3id.org/GDPRov
[13] https://w3id.org/GConsent

concepts.

1.4.2.2 GConsent

The determination of consent validity under the GDPR requires additional information [53], [54] which is not captured using GDPRov as it does not relate to representation of activities and artefacts. Therefore, a separate ontology called GConsent was created (and is the basis for formulating research objective $RO3(c)$) to provide necessary concepts and relationships for representing information relevant for management of consent. GConsent focuses on representation of *only* consent information as required to evaluate its compliance. It acts as a distinct modular ontology which can be used by itself to represent consent, or in conjunction with GDPRov to represent consent and its related activities.

While GDPRov and GConsent both represent consent, the focus of GDPRov is on representing activities and artefacts associated with consent, while GConsent represents information associated with management of consent based on GDPR compliance requirements. Another perspective on this is that GDPRov represents a specific semantic view based on the notion of capturing provenance of activities in ex-ante and ex-post phases, while GConsent represents a state-based representation of consent. The application of these ontologies within use-cases in Chapter 6 show, both ontologies share some concepts and overlap, but are complimentary in their use and represent different aims in their representation of information.

GConsent provides the necessary concepts and relationships to express information about consent in terms of entities such as individuals or agents, purposes and processing, involvement of third parties, medium and context of provision, relationship between instances (e.g. withdraws, updates), and the novel concept of 'consent states' which enables management of consent as an entity. In comparison with state of the art, GConsent provides greater representation of information related to consent and is the most comprehensive ontology for representing consent (based on comparisons demonstrated in Chapter 5).

1.4.3 Querying Information Related to Compliance using SPARQL

A minor contribution of this thesis is the utilisation of SPARQL to query information relevant for GDPR compliance, which fulfils research objective $RO4$. The use of developed ontologies, namely GDPRtEXT, GDPRov, and GConsent - provide representation of concepts associated with GDPR for use in SPARQL queries to represent compliance questions derived from state of the art (see Chapter 4).

Where approaches in state of the art also use SPARQL to represent questions for compliance [40], [55], the work presented in Section 6.1 is the only one within the state of the art to demonstrate derivation of queries from questions associated with an investigation of compliance, i.e. compliance questions as presented in Section 4.2. A practical application of this demonstrates SPARQL queries derived from questions provided by the Irish Data Protection Commission for assisting organisations with their GDPR compliance readiness [56] and shows use of SPARQL in assisting the investigation process associated with compliance. This application of SPARQL was published in a peer-reviewed publication [57] and was presented to members of the Irish Data Protection Commission as part of research developed in this thesis.

1.4.4 Framework for Validating Information using SHACL Compliance

Another contribution of this thesis is the approach for using SHACL to validate information and linking the results with relevant clauses of GDPR for compliance, which fulfils research objective $RO5$. While SPARQL is sufficient to query information and in some cases to determine compliance based on presence or absence of information, the use of SHACL provides a standardised approach for validation of information based on representing constraints and persisting the results of validation.

The validation using SHACL is part of a proposed framework presented in Section 6.2 which consists of creating a 'compliance graph' for storing information relevant in the investigation and demonstration of compliance. The validation requirements are derived from constraints and assumptions based on compliance questions in Section 4.2.3, and are represented using SHACL with a link to relevant clauses of the GDPR defined using GDPRtEXT to indicate their role in the compliance process. The constraints expressed using SHACL utilise concepts and relationships from GDPRov and GConsent to represent validation requirements, and re-use SPARQL queries created for $RO4$ to retrieve information. The validation results are persisted and annotated with GDPRtEXT to link them with the GDPR, thereby providing a form of documentation for information validation associated with compliance.

The framework suggests a more efficient form of validation by reusing ex-ante validation results in ex-post evaluations by abstracting common constraints belonging to ex-ante information and validating them in the ex-ante stage itself so that only specific constraints associated with instances in the ex-post stage - such as provenance information - need to be validated. The demonstration of the framework and approach consists of evaluating consent on a real-world website to generate a 'compliance report' listing status of validations linked to GDPR. The framework and approach have been published in peer-reviewed publications [58]–[60]

Related work in state of the art uses a variety of approaches for validation and assessment of compliance. The SPECIAL project demonstrates use of OWL2 reasoners to validate consent at ex-ante and ex-post stages [52], [61] and the application of ODRL policies as a compliance checking mechanism [40], [62]. The MIREL project proposes the use of deontic logic for legal reasoning using LegalRuleML [55], [63], while the BPR4GDPR project proposes checking provenance logs for conformance to predetermined processes (ex-post analysis) [64]. The use of SHACL utilising P-Plan workflows to validate policies expressed in ODRL for GDPR compliance has been proposed [65] as a doctoral consortium paper - which provides future directions for application of this research. Compared to state of the art, the approach presented in this thesis is novel in its utilisation of SHACL to validate information and link its results with the GDPR for compliance. It is also novel in its combination and reuse of ex-ante and ex-post validations for compliance.

1.4.5 Information Interoperability Model of the GDPR

A minor contribution of this thesis consists of an information interoperability model based on representing categories of entities (stakeholders) as defined by GDPR and their interactions with respect to interoperability of information shaped by GDPR compliance require-

ments. The model, described in Section 4.1, conceptualises interactions between stakeholders based on information identified as part of $RO1$ and $RO2$, and provides an overview of requirements regarding information and interoperability shaped by GDPR.

The model provides categorisation of information requirements based on provenance, agreements, consent, certification, and compliance; and assists in exploration of existing standards - including semantic web - by outlining requirements and applications of information based on interoperability between entities. It advances state of the art by providing the first systemic analysis of information flows and interoperability between stakeholders, and serves to provide a framework for developing and evaluating potential consensus on interoperability of information for compliance between stakeholders. The model, its analysis, and application in the context of right to data portability was published in peer-reviewed publications [66], [67] and as a book chapter [68].

1.4.6 Participation in DPVCG

The Data Privacy Vocabularies and Controls Community Group[14] (DPVCG) is a W3C community group working towards developing a vocabulary associated with personal data processing based on relevant laws such as GDPR. The group was created by members of the SPECIAL project in May 2018 and currently consists of community members from diverse domains such as academia, legal experts, lawyers, and industry stakeholders.

The work done within DPVCG in its 18 months of operation has produced the Data Privacy Vocabulary[15] (DPV) - an ontological resource for representing information associated with processing of data. The DPV represents a community agreement of vocabulary and semantics of terms and concepts associated with GDPR, and provides a degree of interoperability in representing information for legal compliance. The work regarding creation of DPV has been published in a peer-reviewed conference [69], and has also been listed as a deliverable within the SPECIAL project [70]. The author of this thesis is listed as an editor and contributor in both publications, and is the co-chair DPVCG since January 2020.

The research presented in this thesis had an impact in the creation of DPV through use of developed ontologies as an input as well as through direct participation of the author as an active contributing member. An overview of DPV is therefore presented in Section 5.5 along with comparisons to developed ontologies (GDPRtEXT, GDPRov, GConsent) and SotA. To summarise the comparison, DPV provides a high-level abstraction of terms and concepts, whereas the ontologies in this thesis provide representations of information with more granularity and detail - which makes their usage with DPV complimentary rather than contradictory.

1.4.7 Publications

The following peer-reviewed publications present the research in this thesis (grouped by relevance, ordered chronologically reversed):

[14] https://www.w3.org/community/dpvcg/
[15] http://w3.org/ns/dpv

1.4.7.1 Ontologies representing information for GDPR compliance

The following publications are associated with $RO3$ - developing ontologies for representing the concepts and relationships within the GDPR.

1. "**GConsent - A Consent Ontology Based on the GDPR**" [71]
   H. J. Pandit, *C. Debruyne, D. O'Sullivan, and D. Lewis.*
   16th European Semantic Web Conference (ESWC), 2019.

   This publication presents the GConsent ontology for representing information about consent as required by GDPR. GConsent fulfils research objective $RO3(c)$, and provides a detailed representation of consent for information management and documentation. GConsent is described in Section 5.4.
2. "**GDPRtEXT - GDPR as a Linked Data Resource**" [72]
   H. J. Pandit, *K. Fatema, D. O'Sullivan, and D. Lewis.*
   15th European Semantic Web Conference (ESWC), 2018.

   This publication presents the GDPRtEXT resource consisting of a linked data representation of the text of GDPR, and a glossary of its concepts. It also provides a mapping from clauses of the DPD to GDPR based on reuse of compliance methods developed for DPD for GDPR. GDPRtEXT fulfils research objective $RO3(a)$, and is instrumental in providing semantic association between information and GDPR for approaches presented in this thesis. GDPRtEXT is described in Section 5.2.
3. "**Modelling Provenance for GDPR Compliance using Linked Open Data Vocabularies**" [66]
   H. J. Pandit, *and D. Lewis.*
   5th Workshop on Society, Privacy and the Semantic Web - Policy and Technology (PrivOn2017), co-located with the 16th International Semantic Web Conference (ISWC), 2017.

   This publication presents the GDPRov ontology for representing the provenance of personal data and consent for GDPR, and discusses use of its concepts in SPARQL queries for retrieving information associated with compliance. GDPRov fulfils research objective $RO3(b)$, and provides ex-ante and ex-post representations for activities associated with personal data and consent for GDPR. GDPRov is described in Section 5.3.
4. **Compliance through Informed Consent: Semantic Based Consent Permission and Data Management Model** [73]
   K. Fatema, E. Hadziselimovic, ***H. J. Pandit****, C. Debruyne, D. Lewis, and D. O'Sullivan.*
   5th Workshop on Society, Privacy and the Semantic Web - Policy and Technology (PrivOn2017), co-located with the 16th International Semantic Web Conference (ISWC), 2017.

   This publication presents an early (pre-GDPR enforcement) collaboration in developing a preliminary ontology for representing consent and a data management model for GDPR. The early work was crucial towards understanding complexities of consent, and provided valuable feedback towards development of GConsent.
5. "**Linked Data Contracts to Support Data Protection and Data Ethics in the Sharing of Scientific Data**" [74]
   E. Hadziselimovic, K. Fatema, ***H. J. Pandit****, and D. Lewis.*
   Workshop on Enabling Open Semantic Science (SemSci), co-located with the 16th International

Semantic Web Conference (ISWC), 2017.

This publication presents an early collaboration (pre-GDPR) towards developing an ontology for representing data sharing agreements for GDPR by extending the ODRL ontology. The ontology enables representation of obligations associated with propagation of rights between parties that share or exchange data.

1.4.7.2 Querying and validating information for GDPR compliance

The following publications are associated with $RO4$ - querying for information, and $RO5$ - validating information for compliance.

6. "**Test-driven Approach Towards GDPR Compliance**" [60]
   ***H. J. Pandit**, D. O'Sullivan, and D. Lewis.*
   14th International Conference on Semantic Systems (SEMANTiCS), 2019.

   This publication presents implementation of approach for validation of information by utilising the use-case of consent in a real-world website. It utilises SHACL to validate information represented by GDPRov and GConsent, and uses GDPRtEXT to associate constraints and results with GDPR. It also demonstrates use of SPARQL to identify tasks and reports related to compliance by querying validation results. The approach demonstrates usefulness of combining ex-ante and ex-post approaches in terms of efficiency and compliance. This research fulfils research objective $RO5$ and is presented in Section 6.2.
7. "**Queryable Provenance Metadata For GDPR Compliance**" [57]
   ***H. J. Pandit**, D. O'Sullivan, and D. Lewis.*
   14th International Conference on Semantic Systems (SEMANTiCS), 2018.

   This publication presents use of SPARQL queries to represent questions associated with compliance by using GDPRtEXT and GDPRov ontologies. It demonstrates effectiveness of SPARQL in retrieving information for GDPR compliance, and fulfils research objective $RO4$. This work is presented in Section 6.1.
8. " **Exploring GDPR Compliance Over Provenance Graphs Using SHACL**" [59]
   ***H. J. Pandit**, D. O'Sullivan, and D. Lewis.*
   14th International Conference on Semantic Systems (SEMANTiCS) - Posters track, 2018.

   This publication presents an overview of the approach for validating information using SHACL and associating results with specific articles of GDPR. The approach proposes persistence of validation results to create a 'compliance graph' that can be queried and validated for documenting information for compliance. This work is presented in Section 6.2.1.
9. "**Towards Knowledge-based Systems for GDPR Compliance**" [58]
   ***H. J. Pandit**, C. Debruyne, D. O'Sullivan, and D. Lewis.*
   International Workshops on Contextualized Knowledge Graphs (CKG), co-located with 17th International Semantic Web Conference (ISWC), 2018.

   This publication explores creation of a knowledge-based framework based on utilisation of information associated with compliance using semantic web technologies for applications such as creation of reports, documentation, and assessment of compliance for different stakeholders. The approach was used in conjunction with the above

publication in addressing research objective $RO5$.

1.4.7.3 Model for information interoperability based on requirements of GDPR compliance

These publications present a model of interaction between entities as defined by the GDPR, and explore information categories and their interoperability requirements based on existing standards, including those provided by the semantic web. The model provides an overview of information flows between stakeholders, and the role of interoperability in facilitating information for compliance between them. This research is presented in Section 4.1.

10. "**Standardisation, Data Interoperability, and GDPR**" [67]
    H. J. Pandit, *C. Debruyne, D. O'Sullivan, and D. Lewis.*
    Book Chapter in Shaping the Future Through Standardization, 2019
11. "**An Exploration of Data Interoperability for GDPR**" [68]
    H. J. Pandit, *C. Debruyne, D. O'Sullivan, and D. Lewis.*
    International Journal of Standardization Research (IJSR) , Vol. 16 Issue. (1), 2018
12. "**GDPR Data Interoperability Model**" [75]
    H. J. Pandit, *D. O'Sullivan, and D. Lewis.*
    23rd European Academy for Standardisation Annual Standardisation Conference (EURAS), 2018

1.4.7.4 Investigated applications of research - Information Management

The following publications do not directly address the research question, but consist of applying the research presented in this thesis towards processes that assist with the compliance process.

13. "**Towards Generating Policy- Compliant Datasets**" [76]
    C. Debruyne, ***H. J. Pandit***, *D. O'Sullivan, and D. Lewis.*
    13th IEEE International Conference on Semantic Computing (ICSC), 2019.

    This publication presents an approach for generating just-in-time datasets consisting of personal data based on given consent to ensure processes are compliant in their usage of consent.
14. "**GDPR-driven Change Detection in Consent and Activity Metadata**" [46]
    H. J. Pandit, *D. O'Sullivan, and D. Lewis.*
    4th Workshop on Managing the Evolution and Preservation of the Data Web (MEPDaW), co-located with 15th European Semantic Web Conference (ESWC), 2018.

    This publication proposes an approach for detecting changes related to use of personal data and consent in activities by utilising the ex-ante component of P-Plan to represent activities and comparing them using a graph-based algorithm.

1.4.7.5 Investigated Applications of Research - Privacy Policies

The following publications do not directly address the research question, but consist of applying the research presented in this thesis towards privacy policies.

15. "**Extracting Provenance Metadata from Privacy Policies**" [77]
    H. J. Pandit, *D. O'Sullivan, and D. Lewis.*

7th International Provenance & Annotation Workshop (IPAW), par t of Provenance Week, 2018.

This publication discusses use of GDPRov to represent extracted information about activities associated with personal data within a privacy policy.

16. "**An Ontology Design Pattern for Describing Personal Data in Privacy Policies**" [44]
***H. J. Pandit**, D. O'Sullivan, and D. Lewis.*
9th Workshop on Ontology Design and Patterns (WOP), co-located with 17th International Semantic Web Conference (ISWC), 2018.

This publication presents an ontology design pattern that uses GDPRov and GDPRtEXT to represent information about personal data and its processing in a privacy policy.

17. "**Personalised Privacy Policies**" [45]
***H. J. Pandit**, D. O'Sullivan, and D. Lewis.*
4th International Workshop on TEchnical and LEgal aspects of data pRIvacy and SEcurity (TELERISE), co-located with 22nd European Conference on Advances in Databases and Information Systems, 2018.

This publication discusses personalisation of privacy policies by using information about an individual's personal data processing and using GDPRtEXT and GDPRov to annotate it for a machine-readable representation.

1.4.7.6 Data Privacy Vocabulary

The following publication presents work related to creation of the Data Privacy Vocabulary by DPVCG and describes the methodology used with relation to the existing vocabularies - including those presented in this thesis - namely GDPRtEXT, GDPRov, and GConsent. The DPV is described in Section 5.5.

18. "**Creating A Vocabulary for Data Privacy**" [78]
***H. J. Pandit**, A. Polleres, B. Bos, R. Brennan, B. Bruegger, F. J. Ekaputra, J. D. Fernández, R. G. Hamed, E. Kiesling, M. Lizar, E. Schlehahn, S. Steyskal, R. Wenning*
18th International Conference on Ontologies, DataBases, and Applications of Semantics (ODBASE), 2019.

1.5 Thesis Overview

The rest of this thesis is structured as follows:

Chapter 2: Background on GDPR and Semantic Web

This chapter presents a summary of information required to understand the work presented in this thesis. The chapter consists of two sections: the first describes concepts and requirements of GDPR, while the second section describes semantic web technologies with an overview of its standards and vocabularies.

Chapter 3: State of the Art

This chapter reviews existing work and approaches regarding regulatory compliance with a specific focus on those addressing GDPR compliance. The chapter starts by providing an overview of approaches used for legal compliance. It then presents an in-depth review of approaches utilising semantic web technologies to address GDPR compliance requirements, followed by other approaches for GDPR compliance. Approaches which do not directly address the GDPR, but are relevant to legislative compliance and semantic web are also presented. The chapter then presents an analysis of the state of the art and concludes with a discussion on identified gaps and limitations.

Chapter 4: Information Required for GDPR Compliance

This chapter presents information required for GDPR compliance of activities associated with processing of personal data and consent in ex-ante and ex-post phases. The chapter starts by presenting an information model for interoperability of information between stakeholders defined by the GDPR. The model provides an analysis of information interoperability requirements based on requirements of GDPR compliance and the role of existing standards in addressing them. This is followed by expressing information requirements as analytical questions - termed 'compliance questions' - whose answers provide the information necessary to evaluate compliance. The chapter then concludes with identification of constraints and assumptions which can be used to validate information for GDPR compliance.

Chapter 5: Representing Information for GDPR Compliance using Ontologies

This chapter presents the OWL2 ontologies developed to represent information associated processing of personal data and consent for GDPR compliance. The ontologies present concepts for answering compliance queries presented in Chapter 4. The first ontology presented is GDPRtEXT - which provides a method to link information with concepts and clauses of GDPR through a linked data version of its text and a vocabulary of concepts. The second ontology presented is GDPRov - which enables representation of provenance information regarding personal data and consent as models or templates and their executions or activity logs. The third ontology presented is GConsent - which enables representation of information associated with consent. The chapter presents an overview of concepts and relationships for each vocabulary, its relation with GDPR, and the competency questions used to guide its development. The chapter also presents a brief overview of the Data Privacy Vocabulary and its comparison with the other presented ontologies and the SotA.

Chapter 6: Querying and Validating Information for GDPR Compliance

This chapter presents use of SPARQL to express compliance queries using ontologies presented in Chapter 5. The chapter also presents a framework to validate information using SHACL based on constraints identified in Chapter 4. The framework demonstrates use of semantic web technologies in validating information for GDPR compliance by utilising a combination of ex-ante and ex-post validations and linking of results with GDPR for documentation of information for compliance.

Chapter 7: Conclusion

This chapter concludes the thesis with a summary of key findings and outcomes of the presented work. It discusses the extent to which the thesis serves to address the research question(s) and objective(s), and outlines directions for future work in terms of potential applications and extension through related work.

This page intentionally left blank.

2 | Background: GDPR and the Semantic Web

This chapter presents the necessary background information for understanding the research presented in this thesis. The first section (Section 2.1) provides a short introduction to the General Data Protection Regulation (GDPR) in terms of terminology, information requirements for compliance, and sources of additional information regarding GDPR compliance. The second section (Section 2.2) provides an introduction to Semantic Web technologies based on the formulation of the research question and objectives regarding information representation, querying, and validation.

2.1 General Data Protection Regulation (GDPR)

The General Data Protection Regulation (GDPR) [3] is the current data protection law applicable within the European Union (EU) and the European Economic Area (EEA) and regulates use and processing of personal data. It supersedes its predecessor - the Data Protection Directive (DPD) [6] - and provides greater requirements and transparency for compliance, with potentially large and significant amount in fines if organisations are found to have violated its obligations. A significant aspect of GDPR are its principles and rights which are intended to afford greater privacy and control to an individual regarding use of their personal data.

By virtue of being a regulation as opposed to a directive, GDPR is considered enforceable law with local and national data protection laws acting in conjunction rather than replacing it. The GDPR has attracted global attention and scrutiny due to its requirements for compliance and potential fines, as well as for providing rights that enhance privacy. It has influenced other privacy laws across the globe with the California Consumer Protection Act (CCPA) being a recent example [79].

2.1.1 Terminology

The legal terminology utilised in GDPR is intended to clarify the roles, actions, and concepts referred in its obligations. The definition of ***personal data*** (Article 4-1) is based on linking, association, or relevance of any information with an individual - and represents a significant change from its predecessor as well from other laws which rely upon the concept of Personally Identifiable Information (PII). The individual the personal data relates to is termed as ***Data Subject*** (Article 4-1) - a distinct term from other relevant laws that use *individual* or *PII Principal*.

GDPR regulates ***processing*** of personal data (Article 4-1) - which is defined as any action over or utilising personal data as: "collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction;" [3].

A ***Controller*** is defined (Article 4-7) as the entity or organisation that determines ***purpose*** and means of processing of personal data. As such, controller is the primary organisation regarding GDPR compliance and is subject to additional obligations given its role in determining processing of personal data. One or more controllers can act together to determine purpose and processing, in which case they are defined as ***Joint-Controllers*** (Article 26).

A ***Processor*** is defined (Article 4-8) as an entity which processes personal data on specified instructions of a controller (Article 28) and is not allowed to deviate from the instructions or utilise the personal data for other purposes. A processor is also referred to as *(sub-)contractor* in some contexts. A processor may appoint other *sub-processors* in order to carry out processing, where sub-processors are required to follow the same obligations as the processor.

A ***Data Protection Officer (DPO)*** is defined (Article 37) as an individual appointed by a controller or processor to oversee compliance and processing of personal data, monitor internal processors, and collaborate with supervisory authorities as required.

A ***Regulatory or Supervisory Authority or Data Protection Commission*** is a governmental organisation with responsibility to evaluate and enforce GDPR compliance. These bodies are established by national or federal governments and have jurisdiction over their appointed regions. The GDPR specifies responsibilities of such bodies and provides avenues for their co-operation across jurisdictions (Article 56, 60, 62).

GDPR provides several ***rights*** to the data subject (Chapter 3) which are obligated to be provided by controller(s). Since rights are mandatory - their implementation, provision, and exercising must be monitored as part of compliance.

Lawful Basis or ***Legal Basis*** is the provision under which processing of personal data is permitted by GDPR, of which there are six primary ones (Article 6). *Legitimate Interest* refers to legal basis where controller or third party needs personal data in order to provide or carry out its operations and services (Article 6-1f). Other legal basis include the carrying out a contract (Article 6-1b), compliance to legal obligation (Article 6-1c), public interest or as part of official authority (Article 6-1d), and other provisions and specifics introduced by national governments (Article 6-2). ***Consent of the Data Subject*** is the legal basis to be used where other legal basis are not applicable, and which requires consent of data subject. Consent is subject to further obligations and requirements depending on the use-case which determine its validity.

2.1.2 Transparency and Requirements

Controllers are required to provide information to data subjects about processing activities regarding type of processing taking place and who is carrying it out. This information is usually part of 'privacy policy' and contains: (a) identity of controller(s); (b) purpose of processing; (c) legal basis; (d) transfer of data and categories of recipients; (e) data storage periods; and (f) the existence of rights.

In addition to above, a controller is required to monitor and maintain documentation detailing their processing of personal data categories along with particulars of data included within each category. Every personal data being processed must be associated with a source. GDPR specifies certain personal data categories as 'special categories' based on their sensitivity and need for additional measures regarding security and accountability. Such special categories contain additional obligations for compliance which must be enforced if an organisation is using them.

2.1.3 Data Protection Impact Assessment

A ***Privacy Impact Assessment*** or ***Data Protection Impact Assessment*** is an obligation of the controller to conduct and document an impact assessment of data processing before it is executed (Article 35). GDPR lays out conditions under which a DPIA is mandatory to be conducted - which includes processing that results in a high risk to rights and freedoms of data subjects, use of special categories of personal data, automated processing with significant effects, and use of evaluation or scoring.

A DPIA can be conducted for a specific processing activity, for a project, or for the entire organisation - and must be carried out "prior to the processing" (Articles 35-1, 35-10). The aim of DPIA is to identify and mitigate data protection risks, plan for implementation of solutions to those risks, and assess viability of a project at an early stage. Documentation of DPIA process allows demonstrating compliance with GDPR and minimising risks of legal difficulties in carrying out the processing.

Carrying out DPIA requires identification of information flows in terms of collection, storage, usage, sharing, and erasure of personal data within defined processing activities. A DPIA needs to be updated or carried out again if there are changes in processing activities.

2.1.4 Sources of Additional Information

2.1.4.1 Data Protection Authorities

A ***Data Protection Authority (DPA)*** is the authority responsible for upholding EU laws and rights regarding privacy and data protection through enforcement and monitoring of compliance. Its title and reference differs by language and jurisdiction - for example the *Data Protection Commission (DPC)* in Ireland is also sometimes referred to as *Data Protection Commissioner's Office*. A DPA is also referred to as *Regulator* given their task of regulating processing of personal data. In some nations (such as Germany) the data protection authorities are established in federal states or regions rather than a singular nation-wide office. The data protection authorities are also governed by national data protection laws in addition to the EU laws such as GDPR, and are responsible for their upholding as well.

The data protection authorities in each of their jurisdictions have published guidance and documents for assisting organisations and data subjects in understanding the GDPR and its compliance requirements. The authorities in Ireland[1], United Kingdom[2], and France[3] have provided this information in English.

[1] https://www.dataprotection.ie/
[2] https://ico.org.uk/
[3] https://www.cnil.fr/en/home

2.1.4.2 Article 29 Data Protection Working Party and European Data Protection Board (EDPB)

The Article 29 Working Party (Art. 29 WP) was an advisory body to the European Parliament and its bodies, and consisted of representatives from data protection authorities of each EU member state, the European Data Protection Supervisor, and the European Commission. The body was established following the Data Protection Directive (DPD) and was replaced by European Data Protection Board (EDPB) under GDPR. The working party provided expert advice to member states, and published opinions on application of laws affecting right to protection of personal data - including the GDPR. To this end, it published[4] a number of documents clarifying or expressing opinions on interpretation of GDPR.

The EDPB is the European body replacing the Art. 29 WP with the purpose to ensure consistent application of GDPR and to assist in co-operation between various data protection authorities. EDPB is tasked with issuing guidelines, recommendations, and identifying best practices related to interpretation and application of GDPR. It advises the European Commission on matters related to protection of personal data in EU and EEA by adopting consistency in cross-border cases, encouraging development of codes of conduct, establishing certification mechanisms for data protection, and promoting cooperation and effective exchange of information and good practices among national supervisory authorities. The EDPB has published[5] documents for providing information and assistance to individuals, controllers & processors, and regulators.

2.1.4.3 STAR & STAR II

STAR[6] (Support Training Activities on the data protection Reform) was an European project that provided materials to support training of DPAs and DPOs for GDPR. Its resources are published in an open and publicly accessible form[7]. The project has produced a handbook for assisting stakeholders in understanding GDPR and preparing for its compliance. The project also provides an evaluation questionnaire and compliance checklist [80] consisting of a list of questions and criterion to assess preparedness with requirements of GDPR. The resources published by the STAR projects provide documentation regarding GDPR compliance that is adopted by organisations and regulatory authorities.

2.1.4.4 PRIPARE

PRIPARE[8] (PReparing Industry to Privacy-by-design by supporting its Application in REsearch) was an European research project that provided a set of documents regarding privacy engineering covering activities such as privacy risk management, requirement analysis, design strategies, maintenance and compliance. It published the PRIPARE methodology handbook [81] which provides guidelines for privacy and security by design. The handbook incorporates information based on a draft of GDPR[9] It provides foundational methodolo-

[4] https://ec.europa.eu/justice/article-29/documentation/index_en.htm
[5] https://edpb.europa.eu/edpb_en
[6] https://projectstareu.wordpress.com/
[7] http://www.project-star.eu/training-materials
[8] http://pripareproject.eu/
[9] Thehandbookwaspublishedin2015,andincorporatedknowninformationabouttheGDPRuptothatpointintime.

gies and reference models for carrying out privacy analysis, designing and implementing privacy enhancing systems, with templates for impact assessments and conformance.

2.2 Semantic Web Technologies

The term '*Semantic Web*' refers to extension of World Wide Web with machine-readable and interoperable metadata to enable encoding of semantics with data. This is achieved through use of standards developed by World Wide Web Consortium (W3C) to promote common data formats for interoperability and data exchange protocols using the infrastructure of the web.

2.2.1 RDF, RDFS, and OWL

2.2.1.1 Resource Description Framework (RDF)

The representation of semantics starts with specification of facts or 'knowledge' using RDF[10], which provides representation of information as 'triples' whose collective expression can be visualised as a graph. RDF triples are serialised using syntax languages such as XML[11], or the more human-readable Turtle[12], or the web friendly JSON-LD[13]. RDF triples utilise the same identifier system as world wide web and are specified using IRIs[14] - a generalised form of URIs which itself are a generalised form of URLs[15]. A database storing RDF is called a triple-store and enables creation of graphs of RDF data and provides an interface for querying it.

A RDF triple consists of a subject, an object, and a predicate - where predicate describes the relationship from subject to object. An example of RDF triple specified using the Turtle language to indicate chapter number of an entity is expressed as -

```
<http://example.com/ch2> foaf:name "Background"@en .
```

The subject in this triple is the IRI `<http://example.com/ch2>` representing a resource, with the predicate `foaf:name` representing relationship of associating a name as indicated by the string in object field `"Background"@en` with `@en` specifying use of English language.

The subject is written in shorthand notation indicating it starts with the same IRI as the data file it is described in. An IRI can become too long for readability and is usually written using shorthand prefixes. The predicate is a *property* provided by an external ontology called 'Friend of a Friend' (FOAF[16]), which is referenced by its prefix `foaf` as shorthand for its entire IRI. The object is a string in this case, but could itself be another RDF triple or resource - in which case it would be specified as an IRI.

2.2.1.2 RDF Schema (RDFS)

RDFS provides concepts for expressing classes, properties, and data types in order to represent schemas using RDF. It also provides commonly required relationships such as spec-

[10] https://www.w3.org/TR/rdf11-concepts/
[11] https://www.w3.org/TR/rdf-syntax-grammar/
[12] https://www.w3.org/TR/turtle/
[13] https://www.w3.org/TR/json-ld/
[14] https://tools.ietf.org/html/rfc3987
[15] https://www.w3.org/TR/uri-clarification/
[16] http://xmlns.com/foaf/spec/

ifying labels of resource, indicating related resource, or specifying domain and range of properties. The example in code below expands upon the triple described in RDF to express chapter as a class with its number as a property, and specifies human-readable labels for each resource.

```
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix : <http://example.com/> .
:Chapter rdf:type rdfs:Class ;
    rdfs:label "Chapter"@en .
:chnum a rdfs:Property ;
    rdfs:label "number"@en ;
    rdfs:range xsd:integer .
:Ch2 rdf:type :Chapter ;
    foaf:name "Background"@en ;
    :chnum 2^^xsd:integer .
```

2.2.1.3 Web Ontology Language (OWL)

While RDFS enables expression of hierarchies, more formal representations of knowledge modelling and logic require additional constructs. These are represented using OWL[17] as an ontology consisting of set of 'individuals' (also called classes) and a set of 'property assertions' which relate individuals to each other. An ontology may also consist of set of axioms which place constraints on sets of individuals and types of relationships permitted between them. These axioms provide semantics which can be used to infer additional information using semantic reasoners based on information explicitly provided. Knowledge expressed using OWL can be (and generally is) expressed using RDF, which makes it possible to encode and store ontologies in as RDF data.

2.2.2 SPARQL Protocol and RDF Query Language (SPARQL)

SPARQL[18] is a query language for retrieving data expressed using semantics provided by RDF. SPARQL queries information following the RDF specification, and thus specifies queries to act on data expressed as 'subject-predicate-object' triples. SPARQL queries can also retrieve information from traditional (SQL) databases which store data in non-RDF form through mapping[19] which permits its semantics to be applied across a large variety of existing data stores.

A SPARQL endpoint is an interface for access and querying over stored data. Such endpoints can be exposed over the web to provide querying capabilities through the internet, with DBPedia[20] - a semantic web representation of Wikipedia - providing a well-known example.

2.2.3 Shapes Constraint Language (SHACL)

SHACL[21] is the W3C standard for expressing constraints that validate RDF data. The input over which SHACL constraints are validated is called the *data graph*. Constraints in

[17] https://www.w3.org/TR/owl2-overview/
[18] https://www.w3.org/TR/sparql11-query/
[19] https://www.w3.org/2008/07/MappingRules/StemMapping
[20] http://dbpedia.org/sparql
[21] http://dbpedia.org/sparql

SHACL are called 'shapes' based on the notion of checking if data *fits a shape*, with the set of constraints being validated called as *shapes graph*. SHACL constraints are themselves also expressed in RDF which makes it possible to serialise them as a data graph and perform querying over it.

The output of a validation process is a conformance report which uses a validation report vocabulary provided by SHACL and indicates *failing validations* and status of validation of a whole. A *conformance report* provides boolean indication of whether the any of given set of validations have failed or conversely whether all validations have passed.

SHACL core refers to features defined within the SHACL standard specification. Extensions have been developed which provide additional features for expression and validation of constraints. *SHACL-SPARQL* provides use of SPARQL queries to retrieve data failing a given constraint, and is mentioned within SHACL specification. SHACL validations are performed using a 'validator' - an implementation and interpretation of SHACL standard that provides at least the validation features described in SHACL-core.

Shape Expressions language (ShEx)[22] is an alternative to SHACL that provides a similar conceptual language for expressing constraints and validating RDF data. The design of ShEx emphasises human readability inspired from Turtle, whereas SHACL follows the abstract RDF syntax. ShEx and SHACL were aimed to converge into a common standard, but this has not happened to date - and therefore both are maintained as separate technologies. Both SHACL and ShEx have seen significant adoption by the community.

2.2.4 Standardised Ontologies

2.2.4.1 Provenance Ontology (PROV-O)

PROV-O[23] is a standardised ontological representation of the PROV Data Model[24] (PROV-DM) which is the W3C standard for representing provenance information using semantics provided by RDF and OWL. PROV-O provides classes, properties, and restrictions to represent and interchange provenance information within different contexts. It can be specialised to create new classes and properties to model provenance information for different applications and domains.

2.2.4.2 Open Digital Rights Language (ODRL)

ODRL[25] is a policy expression language that provides an information model, vocabulary, and encoding mechanisms for representing statements about usage of content and services as policies. The ODRL Information Model describes underlying concepts, entities, and relationships that form the foundational basis for semantics of ODRL policies. Policies are used to represent permitted and prohibited actions over one or more assets and can also contain obligations required to be met by stakeholders. Policies can also specify constraints - such as temporal or spatial - and duties which are required to be carried out. ODRL conformances is based on evaluating whether a given information representing an use-case or context satisfies all the expressions described in a given policy.

[22] https://www.w3.org/community/shex/
[23] https://www.w3.org/TR/prov-o/
[24] https://www.w3.org/TR/prov-dm/
[25] https://www.w3.org/TR/odrl-model/

This page intentionally left blank.

3 | STATE OF THE ART

This chapter presents the state of the art (SotA) regarding technological approaches for GDPR compliance with a particular focus and emphasis on those that utilise semantic web technologies.

Legal compliance as a field is quite broad due to the extent of laws which can be abstract (or universal) in their application or are enacted for a specific domain. Achieving legal compliance is itself a large field of research, with technical approaches providing a special attraction in all domains due to potential of automation and co-ordination with information management systems. For the purposes of this thesis, the scope and focus is on technical approaches to assist legal compliance process within the domain of GDPR compliance.

Before delving into approaches concerning GDPR, it is imperative to have an understanding of the field concerning technological approaches used in legal compliance. As all approaches for legal compliance share their motivation and aims, understanding their commonality and rationale enables placing the work associated with GDPR compliance within the broader scope of legal compliance. It also provides investigation of avenues where existing approaches developed for other areas of legal compliance could be reused for GDPR, while providing the necessary background for understanding some of the decisions made by approaches concerning GDPR compliance in reusing existing technologies developed for legal compliance.

Surveys analysing approaches over a period of time provide a good overview of the use of technology in addressing legal compliance. One such survey analyses technological approaches for legal compliance within the 50 years of 1957-2007 [12] to provide categorisation based on representation of information and a set of Requirements Engineering Objectives (REO) an approach must follow to assist the legal compliance process. While the survey paper does not explicitly refer to linked data principles [19], some of the REO are fulfilled by utilising linked data and semantic web to provide an interoperable representation based in standards to define the required metadata regarding legislations.

Where legal compliance is closely tied to functioning of an organisation, requirements of compliance need to be incorporated into relevant activities (commonly called *business processes*) within the organisation. As GDPR regulates use of personal data and requires organisations to adopt practices such as provision of rights, the study of compliance approaches from the perspective of business processes is also relevant in understanding state of the art. For this, two key surveys provide the necessary overview of relation between business processes and compliance. The first survey [15] relates to use of business processes within different phases of compliance, and provides basis for categorising approaches as ex-ante (a-priori, ad-hoc) and ex-post (a-posteriori, post-hoc) depending on whether compliance is

relevant before the activity (in this case processing of personal data) has taken place or after. The second survey [16] analyses approaches utilising business processes for compliance and concludes that compliance approaches are basically limited to identifying relevant requirements from laws and regulations and ensuring that business processes comply with them without much attention to compliance enactment.

'Legal Ontologies' is an umbrella term used to represent ontologies that represent information associated with the legal domain. Key surveys in this domain categorise and analyse ontologies based on purposes [25][43], access control [18], rights expression languages [82], privacy policy languages [83], and privacy requirements engineering [84]. Ontologies have been extensively used for formal representations of information, rules [85], rights [82], legislations [43], business processes [17], policies [83], and requirements [84]. In areas of compliance, standardisation of representations for norms and requirements has seen efforts such as LegalRuleML [86] which uses temporal and defeasible logic, and has been applied for compliance checking over business processes [87]. On similar lines, Compliance Management Ontology [88] proposes shared conceptualisation of business process management, culture management, obligations, programme, resources, risk management, and solutions. There has been a noted lack of guidelines regarding their reuse, particularly within the legal compliance domain [89] where evaluation of developed ontologies is a challenge due to requirement of specialists or legal experts being involved in modelling stage [25].

The state of the art in technologies used for legal compliance thus presents encompassing representations for documents, requirements, rules, obligations, policies, and business processes associated with legal compliance. An adopter has a range of approaches to choose from based on existing work and formalisms where the selection of an approach is dependant on requirements, goals, and its role in legal compliance process.

Given the motivation of this thesis in utilising semantic web technologies and the research question exploring activities associated with personal data and processing for GDPR compliance, the state of the art is presented as an overview of approaches within the scope of these topics. The other approaches are also presented to for relevance of work developed within the domain.

The chapter is structured as follows: the first section (Section 3.1) presents an overview of the methodology used to identify and analyse state of the art. The second section (Section 3.2) presents approaches using semantic web technologies to address GDPR compliance, with the third section (Section 3.3) presenting approaches addressing GDPR compliance using technologies other than semantic web. The fourth section (Section 3.4) presents approaches regarding privacy policies in context of GDPR which feature information relevant to compliance process while not directly addressing compliance itself. The fifth section (Section 3.5) presents approaches that concern representation of consent as required by GDPR. The chapter concludes with an analysis of SotA (Section 3.7). It identifies gaps within SotA and discusses research opportunities to address them.

3.1 METHODOLOGY

3.1.1 Identification and classification of approaches

Identification of approaches was carried out throughout development of research presented in this thesis given the evolving nature of information about GDPR and legal opinions on its compliance. This exercise was carried out from March 2016 until September 2019 at which point the analysis presented here was carried out and encoded into the thesis. Given the research question and motivation of this thesis, the scope of work considered as state of the art was defined as: *approaches addressing GDPR compliance through technological solutions* - with a particular emphasis on the approach being published in an accessible manner so as to enable it analysis.

Peer-reviewed publications were the primary source of knowledge regarding approaches, and were identified using scholarly indexing services such as Google Scholar[1], IEEE Explore[2], ACM Digital Library[3], Scopus[4], and DBLP[5]. In addition to these, information was gathered through dissemination networks such as Twitter[6] and public mailing lists. Searches using keywords such as *GDPR*, *GDPR Compliance*, and *Consent* were used to identify relevant approaches in these sources and added to a reference manager (Zotero[7]). Authors and affiliations of identified publications were also used as keywords to find additional relevant resources. In cases where publications acknowledged funding or projects, an effort was made to identify its online website and access the list of publications. This provided information about the project's aims and objectives, and its future goals and directions.

Publications behind pay-walls which could not be accessed either as a pre-print found elsewhere or by requesting authors for access were not included in state of the art and are not presented in this thesis as they cannot be analysed or compared without access to the work. Requests for access to resources mentioned within a publication was made where possible, with a resource considered 'open' if published and 'accessible' if it could be accessed. Commercial solutions of GDPR compliance are not considered to be within state of the art given their closed nature and lack of knowledge regarding implementations.

The identified approaches were collected and categorised based on their relevance to the research question of this thesis. Approaches addressing GDPR with a particular emphasis on utilising semantic web are presented in Section 3.2 with other approaches addressing GDPR compliance presented in Section 3.3.

3.1.2 Criteria for Analysis

Analysis of state of the art is based on two surveys regarding classification of approaches for legal compliance [12] and phases of compliance [15] that provide context and features to represent the analysis. The research questions and objectives provide the basis for constructing a set of questions for investigation of approaches within state of the art. The questions

[1] https://scholar.google.com/
[2] https://ieeexplore.ieee.org/
[3] https://dl.acm.org/
[4] https://www.scopus.com/
[5] https://dblp.uni-trier.de/
[6] https://twitter.com/
[7] https://zotero.org/

are stated below with an indication to relevant research question and objective guiding this thesis:

1. Which aspects of GDPR compliance does the approach target? ($RO1$, $RO2$)
2. How does the approach represent information associated with GDPR and its compliance? ($RO3$)
3. How does the approach link information with GDPR? ($RO3(a)$)
4. How does the approach represent activities associated with processing of personal data and consent? ($RQ3(b)$)
5. Does the approach provide indication of ex-ante and ex-post phases of activities for compliance? ($RO2$, $RO3$)
6. How does the approach represent information associated with consent? ($RO3(c)$)
7. How does the approach query information relevant to compliance? ($RO4$)
8. How does the approach validate information relevant to compliance? ($RO5$)
9. How does the approach evaluate GDPR compliance? ($RO5$)

Based on these, approaches within SotA are analysed to identify if they: (i) represent clauses and concepts of GDPR, (ii) provide a data model of concepts such as an ontology, (iii) represent information for ex-ante compliance, (iv) represent information for ex-post compliance, (v) model process flows/activities, (vi) model consent information, (vii) evaluate GDPR compliance, (viii) specify requirements for compliance, and (ix) provide their resources in an open and accessible manner to enable its analysis.

These questions and criteria are used to assess and analyse approaches with the outcome presented in Section 3.7 at the end of this chapter. The analyses is presented in sections associated with representation of GDPR, representation of activities, representation of consent, querying of information for compliance, and evaluation of GDPR compliance.

3.2 Approaches for GDPR compliance utilising Semantic Web

SPECIAL

SPECIAL[8] (Scalable Policy-aware Linked Data Architecture For Privacy, Transparency and Compliance) is an European H2020 Project that aims to provide technical solutions for data protection requirements associated with use-cases involving big data. It features extensive use of semantic web technologies across a variety of compliance related tasks such as recording provenance logs of data processing [85], providing visual interfaces for consent [90], [91], compliance checking of processing activities based on given consent [62], [92], [93], and efforts towards standardisation of related vocabularies[9] [78], [94]. Collaborations between SPECIAL and other projects have produced relevant work such as an approach for compliance checking using ODRL [40], [62] and the application of SPECIAL's framework in a Smart City use-case [93]. Information about methodologies used, developed technologies,

[8] https://www.specialprivacy.eu/

[9] The SPECIAL project was the primary driving force behind the creation of the W3C Data Privacy Vocabularies and Controls Community Group (DPVCG), of which the author was an active participating member. Information about this has been listed in Section 1.4.6. The Data Privacy Vocabulary (DPV), produced by the DPVCG, has also been documented as a deliverable of the project in D6.5 [70].

and their evaluations is accessible through peer-reviewed publications and public deliverables[10].

SPECIAL uses a distributed ledger to store processing logs for data processing activities that is evaluated for compliance based on given consent in ex-ante and ex-post phases. The logs are stored using Policy Log vocabulary [51] with data processing activities represented by Usage vocabulary [95]. Compliance evaluation is performed using custom reasoning algorithms [61], [96] in a semantic reasoner. Consent request and management is provided through interactive visual interfaces in a web browser [90], [91]. The project has been successfully applied to use-cases provided by its commercial partners[11] as well as in an external use-case related to use of IoT in Smart Cities [93]. Figure 3.1 presents an overview of SPECIAL architecture along with its components and utilised technologies.

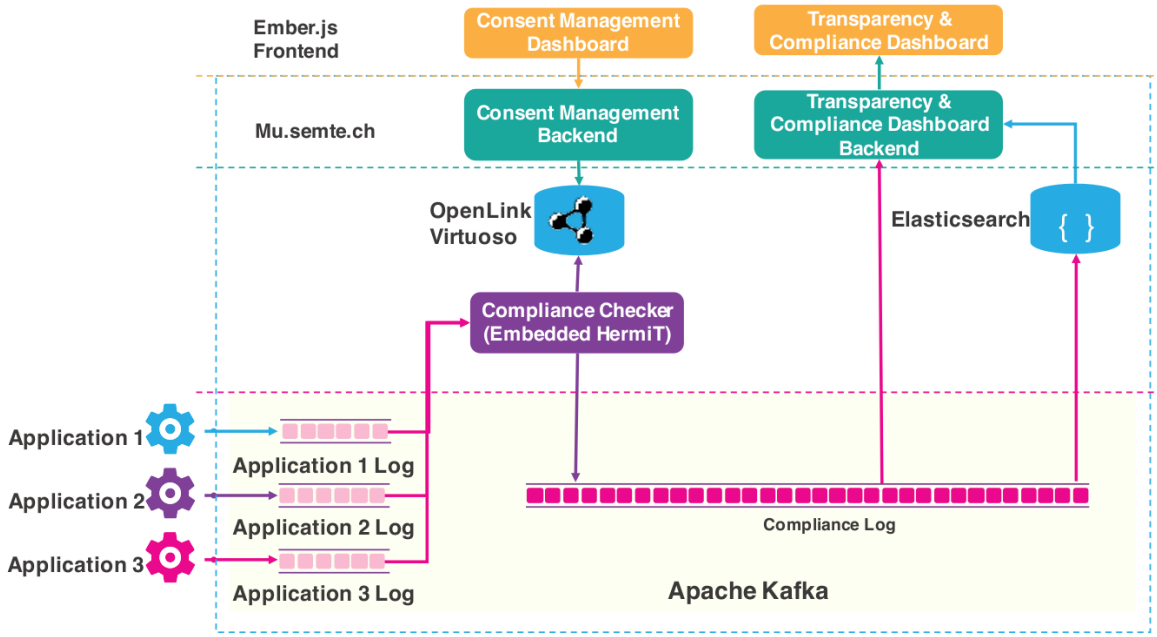


Figure 3.1: Overview of SPECIAL Architecture [85]

The ontology developed to represent data processing activities is called 'Usage Policy' [95] and is described in more detail in D2.5 [99]. In this, 'base policy' is an OWL2 expression denoting intersection of personal data categories, processing operations, purposes, recipients, and storage. Policies are expressed by through OWL2 unions over multiple base policies with differing expressions. SPECIAL provides additional vocabularies for purposes, personal data categories, processing operations, and recipients for using with the Usage Policy based on its use-cases. The general form of a base policy is: The storage expression is itself a policy comprising of OWL2 expressions specifying storage location, duration, and interval. The general form of a storage policy is:

For a given policy P_c representing activities permitted by given consent and policy P_s representing processing activity, compliance is evaluated by checking whether P_c complies with P_s - checked using a semantic reasoner to evaluate $P_c \subseteq P_s$ as an OWL2 subsumption. In SPECIAL's implementation, this is performed using an algorithm [61], [96] implemented

[10] https://www.specialprivacy.eu/publications/public-deliverables
[11] See D1.5 [97] and D1.6 [98] for information on use-cases.

```
ObjectIntersectionOf(
    ObjectSomeValuesFrom(spl:hasData SomeDataCategory)
    ObjectSomeValuesFrom(spl:hasProcessing SomeProcessing)
    ObjectSomeValuesFrom(spl:hasPurpose SomePurpose)
    ObjectSomeValuesFrom(spl:hasRecipient SomeRecipient)
    ObjectSomeValuesFrom(spl:hasStorage SomeStorage)
)
```

```
ObjectIntersectionOf(
    ObjectSomeValuesFrom(spl:hasLocation SomeLocation)
    ObjectSomeValuesFrom(spl:hasDuration SomeDuration)
    DataSomeValuesFrom(spl:durationInDays Interval)
)
```

as a custom semantic reasoner optimised for speed by performing only those operations associated with checking compliance between policies. An extension of the Usage Policy for representing additional requirements of GDPR was discussed in the deliverable D2.6 [100].

Activities are recorded as a log entry using Policy Log vocabulary [51] as represented in Figure 3.2 by `LogEntry`, and are evaluated for compliance using semantic reasoner in both ex-ante and ex-post phases.

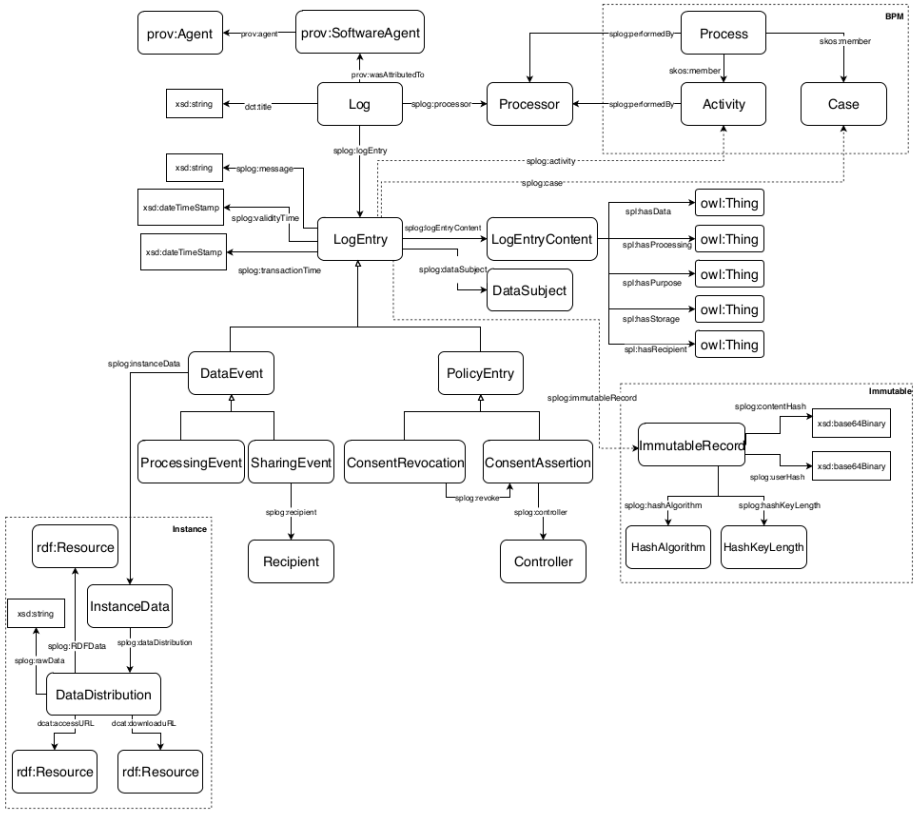


Figure 3.2: Overview of SPECIAL Policy Log Vocabulary [51]

The Policy Log vocabulary, described in D2.7 [101], reuses existing vocabularies of DCAT and PROV-O [47], and is used to record immutable record of activities and their legal basis in consent. Each log entry (`Log`) contains information about agents that created the log

(`prov:SoftwareAgent`), link to data subject, and information about processing through `LogEntryContent` which contains information represented using concepts from base policy. Information about consent is recorded through `PolicyEntry` and consists of consent assertion (given consent) and consent revocation (withdrawn consent).

Methodologies associated with vocabularies produced by SPECIAL project are based on utilisation of metadata in compliance process [102]. The creation of policy vocabularies from use-cases is described through public deliverable D1.5 [97], with creation of log vocabulary described in D3.2 [101]. The vocabularies publicly accessible online with their documentation.

SERAMIS

SPECIAL has produced ODRL models for deontic representations of GDPR requirements through collaborations with other projects. The first in these collaborations is with the Austrian SERAMIS[12] (Sensor-Enabled Real-World Awareness for Management Information Systems) project and has produced a web-based tool called 'PriWUcy' for evaluating compliance assessments of GDPR articles [40], [103]. The tool uses a model of GDPR created representing GDPR obligations using ODRL. A set of preliminary assessment questions provides inputs over which the model is applied to identify actions and obligations to be fulfilled as part of assessment. The developed ODRL model, depicted in Figure 3.3, is extended from the core ODRL model to represent additional constraints of - Feature, Discretional, and Dispensation. The ODRL policies are associated with relevant clauses of GDPR through developed classes representing chapters, articles, and paragraphs within the text of the GDPR. The approach relies on identifying and representing assets, parties, actions, duties, and constraints in GDPR using the developed ODRL model.

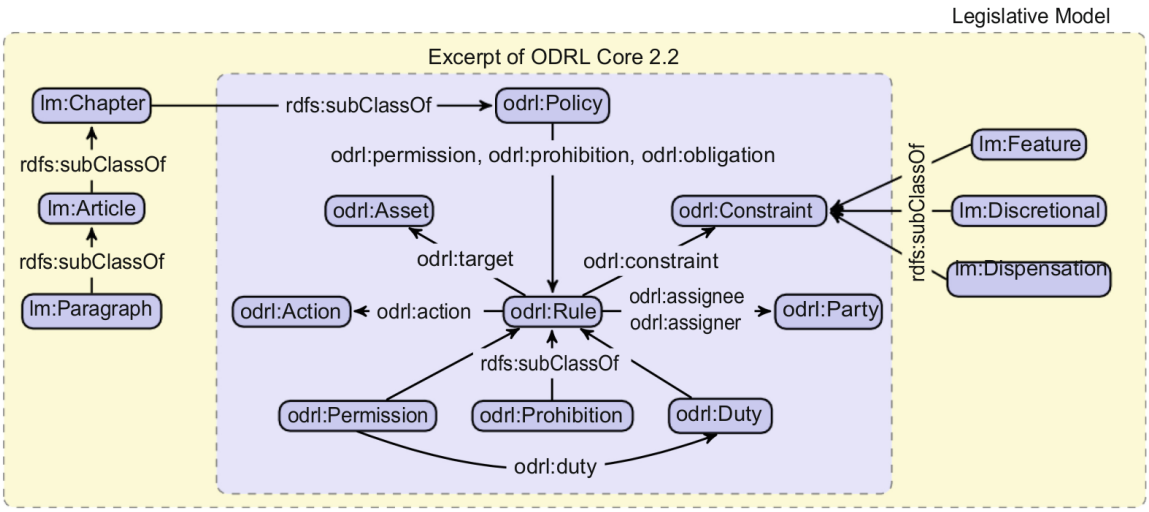


Figure 3.3: Extension of ODRL for representing GDPR obligations [40]

[12] https://cordis.europa.eu/project/rcn/189040/factsheet/en

Vos et al.

The ODRL profile was further developed for modelling regulatory obligations and business policies based on GDPR towards compliance checking and providing explanations of the reasoning process [62]. Additional classes and properties were added for specifying the legal basis, purpose, location, and safeguards associated with processing. Policies representing articles in GDPR are checked against permissions by activities to perform a data processing operation. For example, the following listing shows a request to transfer personal data to a third country based on consent, which would be checked against corresponding policy representing Article 46 of GDPR. Compliance is checked by converting ODRL poli-

```
<http://example.com/policy:bp-transfer> a orcp:Set ;
    odrl:profile <http://example.com/odrl:profile:regulatory-compliance> ;
    orcp:permission
    [ odrl:action orcp:Transfer ;
        orcp:data orcp:PersonalData ;
        orcp:responsibleParty orcp:Controller ;
        orcp:organisationType orcp:InternationalOrganisation ;
        orcp:sender <http://example.com/CompanyA_Ireland> ;
        orcp:recipient <http://example.com/CompanyA_US> ;
        orcp:recipientLocation orcp:ThirdCountry ;
        orcp:purpose orcp:PersonalRecommendations ;
        orcp:legalBasis orcp:Consent ;
        odrl:dataSubjectProvisions orcp:EnforceableDataSubjectRights ;
        odrl:dataSubjectProvisions orcp:LegalRemediesForDataSubjects
    ] .
```

cies into InstAL - a domain-specific language for writing models based on events and states which translates to Answer Set Programming (ASP) and is evaluated using a Answer Set Solver[13]. When a policy is found to be non-compliant, explanations are generated based on 'fluents' - facts that are true if present and false if absent - through a reasoning process. The implementation is publicly accessible through an online repository[14].

CitySPIN

The second collaboration consists of applying SPECIAL's GDPR compliance framework in the Austrian CitySPIN[15] (Cyber-Physical Social Systems for City-wide Infrastructures) project [93] which utilises linked data platforms to ingest data from multiple sources in a smart city scenario and utilises it for data processing pipelines and analytics. An important component of the project is management of consent utilising SPECIAL's policies [51], [95] and compliance framework [85]. Figure 3.4 presents an overview of CitySPIN and its relation with SPECIAL vocabularies. The auxiliary vocabularies provided by SPECIAL for purpose and data categories (SPECIAL-MCM in figure) were extended for use-cases associated with CitySPIN (CPSS Core vocabulary and UC vocabulary in figure). The figure uses white blocks indicating concepts from SPECIAL's vocabularies and greyed blocks indicating additional concepts defined by CitySPIN. The defined policies are checked in ex-ante and ex-post phases using SPECIAL compliance checking algorithm [61], [96], with a proto-

[13] CLINGO https://potassco.org/clingo/
[14] https://github.com/instsuite/instsuite.github.io/blob/master/gdpr.ial
[15] http://cityspin.net/

type implementation demonstrating ex-ante compliance checking in ad-hoc widgets sharing personal data [104]. Details about the project and its implementation are available through public deliverables D6.1 [105] and D6.3 [106].

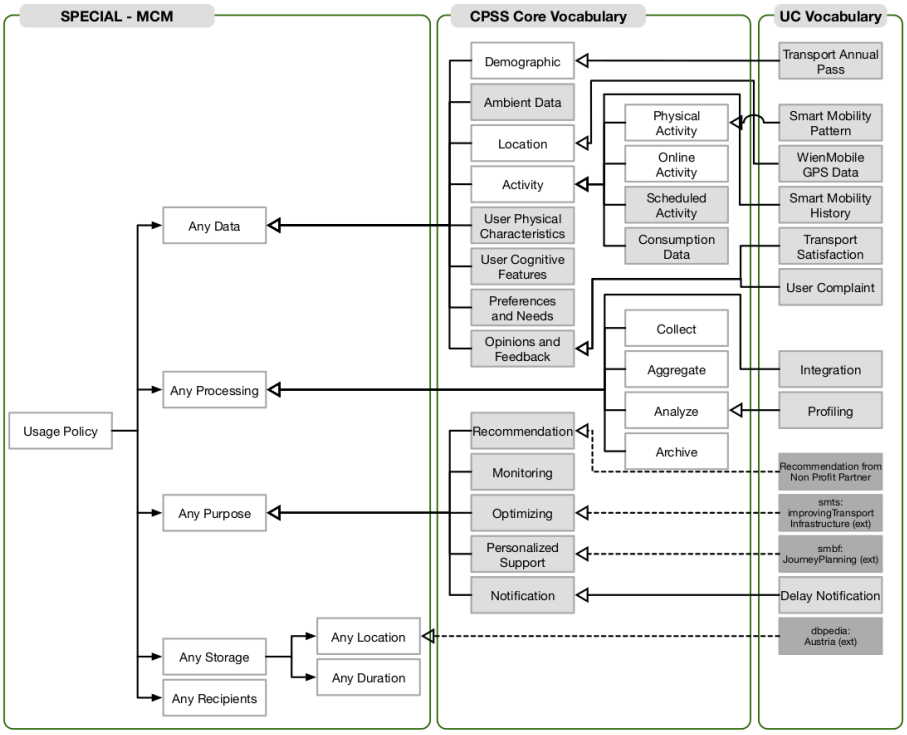


Figure 3.4: CitySPIN's extension of SPECIAL vocabularies [93]

The use-cases utilised within SPECIAL project and a discussion of their legal requirements is provided in deliverables D1.5 [97] and D1.6 [98]. A dashboard representing information of processing activities and consent was created to represent logged information [107] with its usability and testing described in deliverables D4.3 [108] and D4.4 [109]. A separate visual interface for interactive consent was developed based on presenting components of usage policy to individuals in an interactive format [91]. The interface provides visual representations of associations between purposes, processing, data categories, recipients, and storage which allows individuals to explore and manage their consent. It also offers visual paradigms for representation of information as tabs, breadcrumbs, hierarchies, and graphs for visualisation and exploration of information.

The use-cases utilised within SPECIAL project and a discussion of their legal requirements is provided in deliverables D1.5 [97] and D1.6 [98]. A dashboard representing information of processing activities and consent was created to represent logged information [107] with its usability and testing described in deliverables D4.3 [108] and D4.4 [109]. A separate visual interface for interactive consent was developed based on presenting components of usage policy to individuals in an interactive format [91]. The interface provides visual representations of associations between purposes, processing, data categories, recipients, and storage which allows individuals to explore and manage their consent. It also offers visual paradigms for representation of information as tabs, breadcrumbs, hierarchies, and graphs for visualisation and exploration of information.

MIREL

MIREL[16] (Mining and Reasoning with Legal Texts) is an European H2020 Project that aims to create a framework for interpretation of legal texts into formal representations for querying norms, checking legal compliance, and decision support. The project uses semantic web to provide normative requirements [110], use natural language processing over legal texts [111], and creating ontologies based on GDPR [112] for legal compliance [41] and reasoning [55]. The project also involves creation of icons for data protection based on the developed ontologies [113]. Information about the project is available through peer-reviewed publications and public deliverables[17]. Details about the project and its developed resources, namely PrOnto, have been described in publications, but are not open or publicly accessible as of September 2020.

The developed ontology is called PrOnto (Privacy Ontology) [112], and provides concepts regarding GDPR associated with data types and documents, agents and roles, processing purposes, legal bases, processing operations, and deontic operations for modelling rights and duties. It reuses existing vocabularies [55] and has been applied within the Cloud4EU project[18] for legal compliance checking of eGovernment systems as well as the DAPRECO project. PrOnto was developed using MeLOn (Methodology for building Legal Ontologies) developed by the MIREL project [41], [55], [63].

PrOnto consists of modules representing (i) documents and data (depicted in Figure 3.5), (ii) actors and roles, (iii) processing and workflow, (iv) legal rules and deontic formula, (v) purposes and legal bases. It also includes modules for risk analysis and measures - which it utilises to represent risk management processes such as DPIA as workflows [55]. Reasoning is utilised based on the deontic operators within the ontology, and violations are connected with violated obligations thereby providing traceability in steps that created the violation. LegalRuleML is extended to represent obligations and rules for compliance.

A proof-of-concept application for detecting violations of the GDPR [63] utilised PrOnto to model the legal concepts, along with Akoma Ntoso to model the legal text, LegalRuleML to model norms, and Regorous to apply LegalRuleML rules over BPMN and generate a report. The application utilises a web editor for modelling legal rules in connection with the legal text and ontology.

DAPRECO

DAPRECO[19] (Data Protection Regulation Compliance) is an Luxembourgian project relating to the creation of a knowledge base for formal compliance with the terms and provisions of the GDPR. It aims to provide formalisms in deontic logic and natural language semantics for handling legal norms in written language. The project has researched correlation of standards with laws [114], creation of an ontology to model concepts in the GDPR [115], modelling the norms of the GDPR using logic formalisms [116], and annotating BPMN for GDPR compliance [117]. The project involves collaborations with the MIREL project, particularly in the creation and utilisation of PrOnto for addressing GDPR [41], [55], [112], [117].

[16] www.mirelproject.eu/
[17] http://www.mirelproject.eu/publications.php
[18] http://www.agid.gov.it/cloudforeurope
[19] https://www.fnr.lu/projects/data-protection-regulation-compliance/

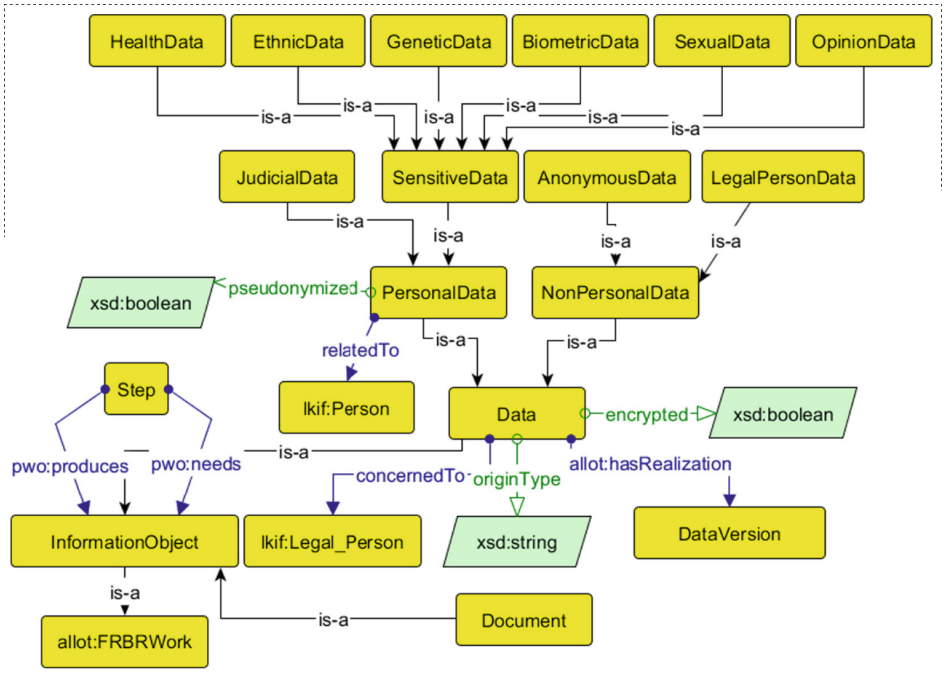


Figure 3.5: Data Categories represented in PrOnto [55]

In turn, DAPRECO provides a practical application of PrOnto as an use-case for legal compliance over business activities [117], [118].

While PrOnto represents a more mature and progressive deliverable of the project's ontology for GDPR, the earlier ontology - referred to as data protection ontology [115] - is relevant for its modelling of concepts and norms based on an draft of GDPR. This ontology uses OWL2 to model concepts of GDPR and was specified to be a work in progress in its publication[20]. Its methodology consists of using structure provided by the Handbook on European Data Protection Law[21], and consists of concepts associated with - data protection principles, rules of data processing constituting duties of data controller, and data subject's rights. The ontology reuses concepts from LKIF Core [119] to represent rights, rules, legal and natural persons. It establishes relations such as *hasObligation* between a *Controller* and a *LegalObligation* concept, and *consentGrantedBy* between *Consent* and *DataSubject* - representing relations derived from GDPR. The ontology was evaluated using OOPS! [38] and was documented using LODE [120]. The ontology can be accessed publicly[22]. The Competency Questions (CQ) representing functional requirements for the ontology were used in its evaluation through SPARQL queries, and are follows:

1. What are the obligations of a data controller?
2. What are the functions of a data processor?
3. What are the rights of the data subject?
4. How do the rights of the data subject relate to the obligations of the data controller and the functions of the processor?
5. How can a data subject interact and/or enforce their rights against a data controller?

[20] It can be assumed the ontology is no longer being used or developed and was utilised in the development of PrOnto based on subsequent publications of the project [116], [117].

[21] 2014 edition https://publications.europa.eu/en/publication-detail/-/publication/eee277d3-d26f-4a18-825b-8e1fd80d2f0d/language-en

[22] Link from [115] https://github.com/guerret/lu.uni.eclipse.bpmn2

6. What are the possible fines and sanctions issued in response to violations by data controllers?
7. Who supervises a data controller?

Subsequent applications of PrOnto in the project utilise Reified Input/Output logic (RIO) [121] - a formalism for normative reasoning based on reification applied to natural language semantics. The rules are expressed using LegalRuleML and are associated with activities using BPMN. This is utilised to allocate tasks related to data protection to stakeholders, and to track compliance across different phases of a software development life-cycle. The project uses these to create a knowledge base [118] with feedback from stakeholders by providing a human-readable version of RIO logic rules utilised to model the GDPR. The project also researchers automating comparison of GDPR with ISO/IEC standards by extracting RIO rules from their texts and analysing them for comparisons. The resulting knowledge base is used in collaboration with MIREL to provide semantic extraction and annotation over ECHR (European Court of Human Rights) judgements [122] as described in MIREL's deliverables D2.4 [123].

BPR4GDPR

BPR4GDPR[23] (Business Process Re-engineering and functional toolkit for GDPR compliance) is an European H2020 project that aims to provide a reference compliance framework for GDPR. The framework is stated to provide specification of sophisticated security and privacy policies, modelling technologies and tools for incorporation of provisions in process models and resulting executable processes, with means for automating verification and alignment. This is planned to be achieved using mechanisms for automating procedures that result in processes that are compliant by design. The project intends to implement the concept of Compliance-as-a-Service (CaaS) by providing a set of tools that fit the needs of various organisations subject to GDPR compliance. Information about the project is available through peer-reviewed publications and project deliverables[24].

The project uses an ontology to represent the information model as depicted in Figure 3.6 [124]. The ontology consists of concepts and relationships regarding data types, roles, operations, and organisation types amongst other contextual concepts. The ontology is serialised using OWL2 and contains 'default instances' representing a ready to use set of instances based on use-cases described in deliverable D3.1 [124].

Compliance assessment is performed using the compliance meta-model ontology depicted in Figure 3.7 and described in deliverable D2.3 [125]. The compliance ontology is used to dictate and evaluate processes by considering them as workflows where actions or operations are connected to each other in terms of dependencies and data flows performed by actors which can include assets or artefacts.

Process mining is performed on knowledge extracted from event logs of information systems to discover, monitor and improve processes not assumed or modelled prior to evaluation. This is utilised to create a process monitoring architecture which contains four subcomponents of: functionality associated with pre-processing, conformance checking, rules, and model repair. The approach intends to 'repair' process models by analysing failed con-

[23] http://www.bpr4gdpr.eu/
[24] http://www.bpr4gdpr.eu/results/deliverables/

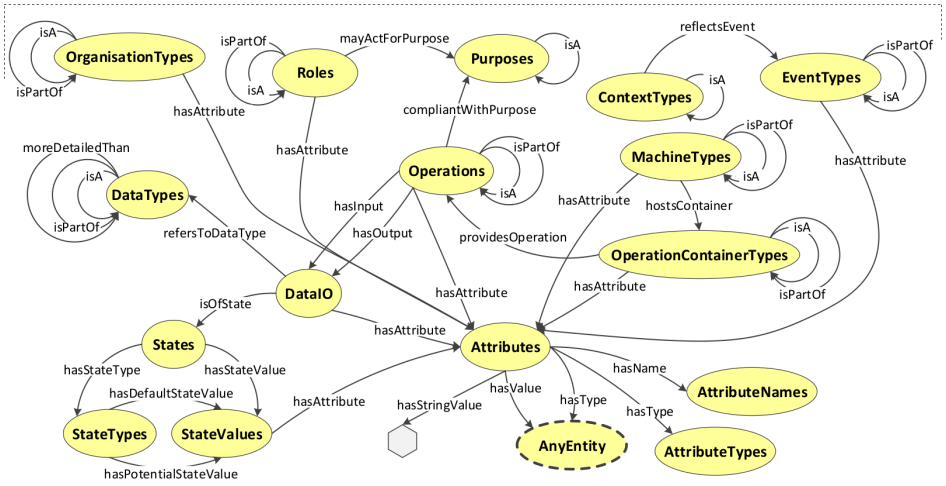


Figure 3.6: Information Model Ontology in BPR4GDPR [124]

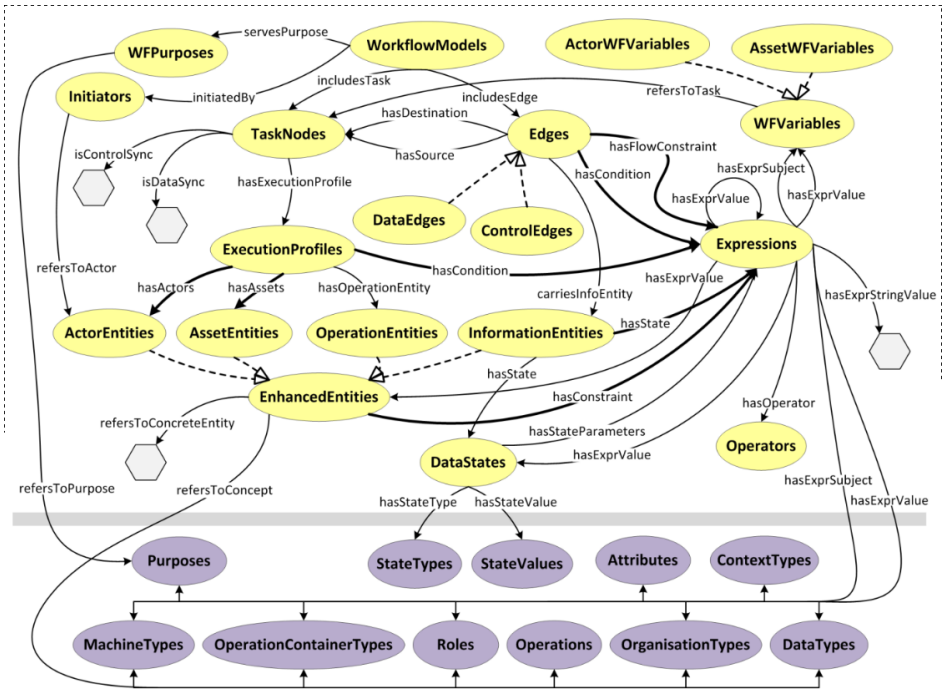


Figure 3.7: Compliance meta-model ontology in BPR4GDPR [125]

formance checks, converted rules, and results of conformance checking to identify parts of process model not compliant with rules.

The public deliverable D3.1 [124] defines rules derived from GDPR which represent a minimal set of configurations for BPR4GDPR architecture before it is adapted to a particular use-case, i.e. its default configuration. These rules are intended to act as constraints in conformance checking and used to repair the processes by identifying components that need to be changed to satisfy rules.

Elluri et al.

Elluri et al. [126], [127] present an ontology of rules and obligations for cloud data providers and consumers based on GDPR for Payment Card Industry Data Security Standard (PCI DSS). The ontology for GDPR [126] is comprised of concepts for components - 'Consumers and Providers', 'Fines and Enforcement', 'Breach & Notification', 'Data Protection Officer', and 'Data Subject'. This is combined with ontology developed for PCI DSS and Clous Security Alliance (CSA) for addressing their combined legal requirements and obligations [127]. The resulting ontology, depicted in Figure 3.8, consists of concepts and relationships associated with stakeholders, obligations for providers and consumers, and security requirements.

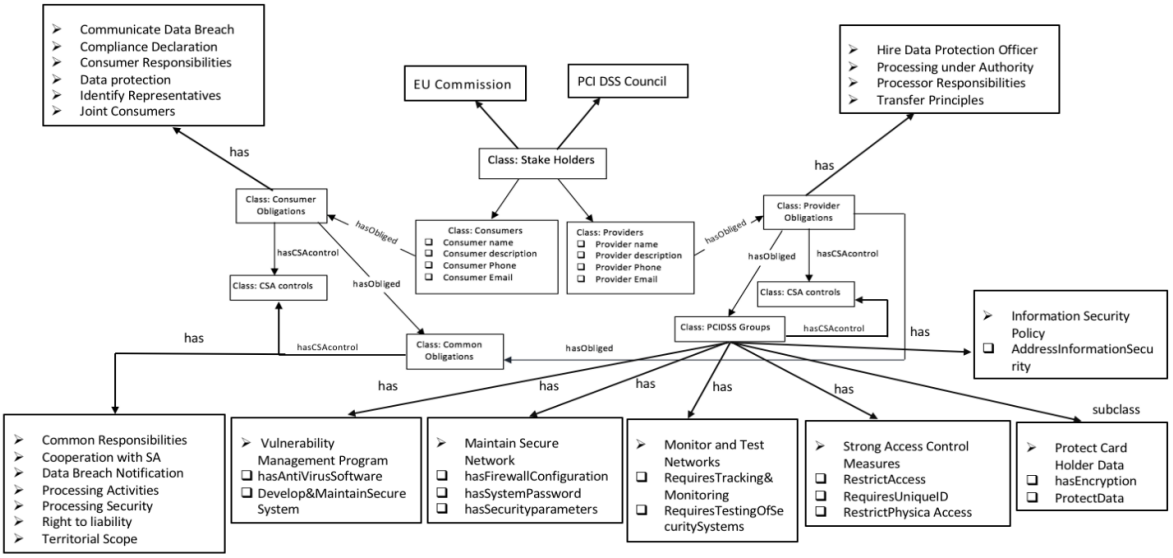


Figure 3.8: Ontology of GDPR, PCI DSS, and CSA by Elluri et al. [127]

The development methodology used to develop the ontology had three steps: (i) Preprocessing, (ii) Ontology Development, and (iii) Validation. The preprocessing stage consisted of extracting relevant chapters and concepts from text of GDPR, PCI DSS, and CSA - and representing them using a preliminary ontology. These were then mapped with corresponding terms from CSA ontology. This stage also involved identifying data protection rules as permissions, obligations, dispensations, and prohibitions using a keyword-based approach. Matching of concepts and rules was performed using a tool that utilised text extraction and regular expressions to match contents with developed ontologies.

The validation of ontology compared text of privacy policies of cloud service providers

using the developed ontology and tools. The extracted terms from policies were populated as instances of concepts from developed ontology to create a RDF representation of policy. Apart from this description, the papers do not provide further details of semantic representations or examples of extracted information, nor an use-case demonstrating capabilities. The ontology is referenced with an accessible link[25] [126], though it does not provide documentation or examples of its usage. Analysing the ontology shows instances of relevant chapters and articles within GDPR applicable to PCI DSS and CSA, and declared as generic resources without reference to actual text or IRI of GDPR.

Related work within the same project by Joshi and Banerjee [128] concerns use of ontology to represent personal data categorisations and data privacy policies for automated enforcing of access control mechanisms regarding sharing of data with third parties through a blockchain. The data privacy policy, depicted in Figure 3.9, shows concepts and relationships for information regarding processing. The privacy policy contains information about data collection purpose, data collected, protection measures, use limitations, consent requirements, and access control.

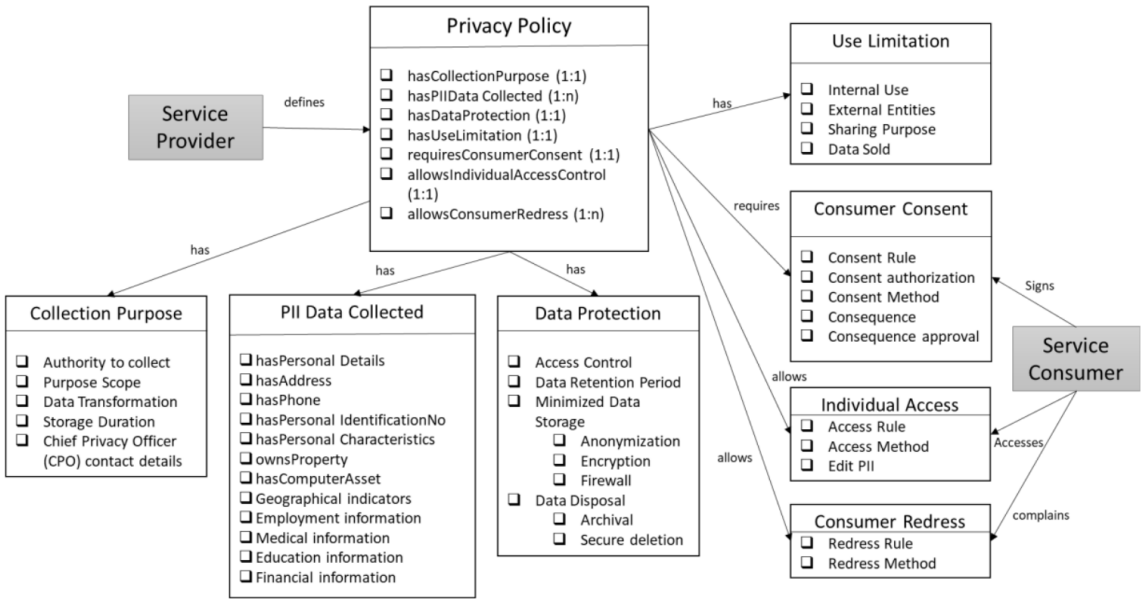


Figure 3.9: Data Privacy Policy by Joshi and Banerjee [128]

The modelling of concepts includes some notably different relationships such as association of storage duration within collection purpose, usage limitation relating to data recipients, attributes associated with consent, and inclusion of redress. The developed ontology is utilised to extract and represent instances from privacy policies, enhance it using a reasoner, and to store resulting access control representations in blockchain. The work is based on utilising an existing privacy policy as a smart contract for sharing of data between provider and consumer. It is important to note that the work relates to use of PII as opposed to personal data as defined by GDPR. Furthermore, the policy does not encode information about rights as required by GDPR.

[25] Ontology by Elluri et al. https://ebiquity.umbc.edu/resource/html/id/377/Ontology-for-EU-s-General-Data-Protection-Regulation-GDPR

Ujcich et al.

Ujcich et al [50] present a data provenance model of GDPR utilising PROV-O [47] to represent activities and data as semantic-web representations. The data model from the publication is depicted in Figure 3.10 and provides an overview of concepts using PROV-O [47] diagrams to denote concepts. The approach uses the activity '*Justify*' to represent rationale for processing personal data and has annotations depicting its starting and ending times, as well as a '*Justification*' - which represents legal basis of processing. The ontology focuses on workflows involving exercising of rights under GDPR, and provides examples of provenance for data rectification and withdrawal of consent.

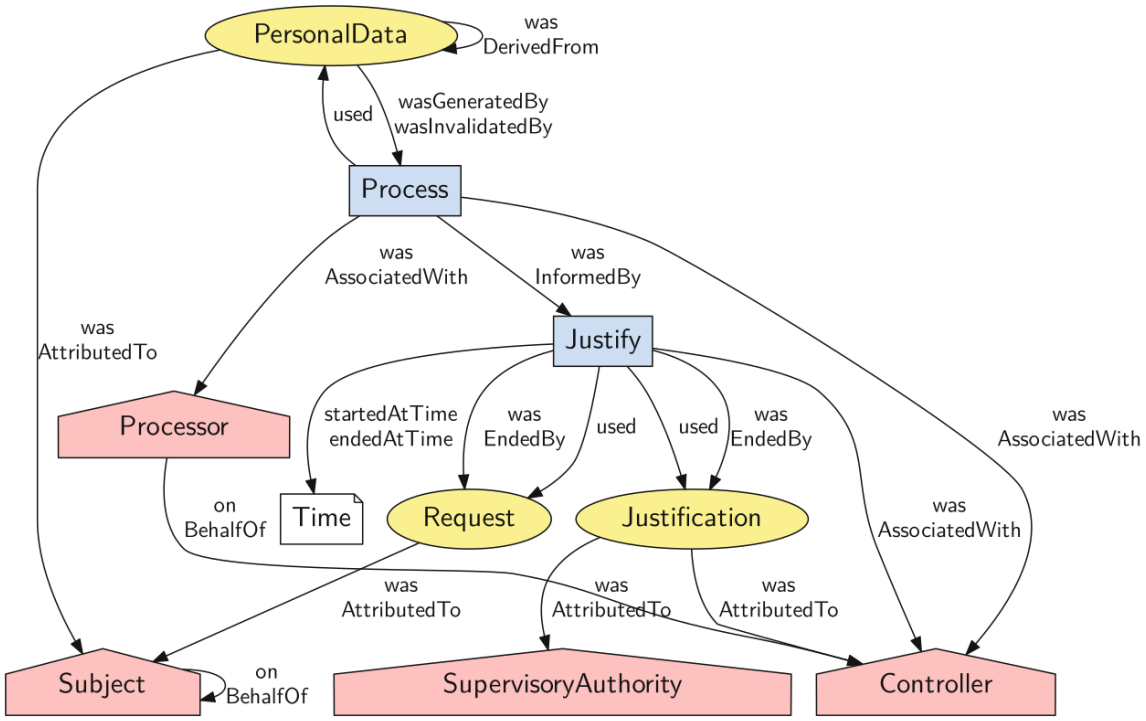


Figure 3.10: GDPR Data Provenance Model by Ujcich et al. [50]

The approach presented in the publication uses design patterns to describe use of ontology in specific cases regarding exercising of rights. The design pattern for provenance associated with consent does not associate data directly with consent owing to data being potentially updated independent of consent. The design pattern is further extended in the publication to represent data sharing with a processor for marketing. The publication provides a discussion on verification of compliance based on provenance through questions which can be queried over the data represented using presented ontology.

Specifying purposes as defined by GDPR is not clarified in this approach as the concept of *Justify* is temporal in its use and limited to processing of personal data it is associated with. The approach also does not provide any information on modelling of more complex processing operations where more than one legal basis could be applied, or where processing operations need to be expressed in several steps. The approach also does not model information about consent or rights, and does not provide ex-ante representations of activities.

Personal Information Controller Service (PICS)

PICS is a project with the objective of enabling users to discover and control use of their personal data by providing a platform for authentication and management between service providers and personal information. The platform architecture [129] consists of three components: (i) personal data controller - provides an user interface allowing users to connect and mediate data to third-party services, (ii) ToS analyser - interprets terms of service and policies of providers using an ontology, and (iii) data mining traces of use - which enables discovery of web services utilising personal data shared with service provider. The platform delegates data storage to an external storage and does not store personal data itself. Authentication is performed at every interaction, and requires users to provide physical authentication using a wearable device.

The ontology used by ToS analyser [130] consists of concepts associated actions on data (such as deletion, disclosure, fusion, sharing), governance (laws, ownership, rights, terms of service), security strategies, and data protection. The ontology itself is not public or accessible and the publication does not provide details about its concepts or use beyond that of utilisation in ToS analyser.

ADvoCATE

ADvoCATE [131] is a consent management platform for data processing in IoT that uses blockchain to record given consent, detects conflicts in data sharing policies, and provides recommendations for managing unwarranted exposure of personal data. In ADvoCATE, the consent notary component is responsible for obtaining and verifying consent signed by both the controller and data subject, and is used to prepare smart contracts for processing based on given consent. The consent management component utilises smart contracts to update user policies to enforce access control over stored personal data. Blockchain is utilised to store versioned hashes of consent linked to smart contracts enforcing them to provide traceability of consent by maintaining a record of consent and its relevant actions.

ADvoCATE has two components - Intelligent Policies Analysis Mechanism (IPAM) and Intelligent Recommendation Mechanism (IReMe) - to identify contradictory rules and policies regarding consent and adapt privacy strategies in real-time. A reference implementation described in the publication uses ontology of GDPR concepts by Bartolini et. al [132] to model data protection requirements, and uses Ethereum as a blockchain platform with Solidity for programming smart contracts. Examples of consent requests presented contain information about personal data category with collection medium and period, retention period, purpose, recipients (EU/Non-EU), use of automated processing, profiling, manual processing, and IDs for controller, device, and device type.

Ontology by Geko & Tjoa

Geko & Tjoa [133] proposed an ontology for capturing interdependence between GDPR and information security requirements for assisting with compliance requirements involving security. The ontology, depicted in Figure 3.11, consists of hierarchical concepts representing obligations of GDPR and information security. It has concepts regarding entities (Controller, Processor, Data Subject), rights, principles, personal data, obligations, and security. These

are related using properties associated with fulfilling roles and obligations, ensuring data security, and provision of rights.

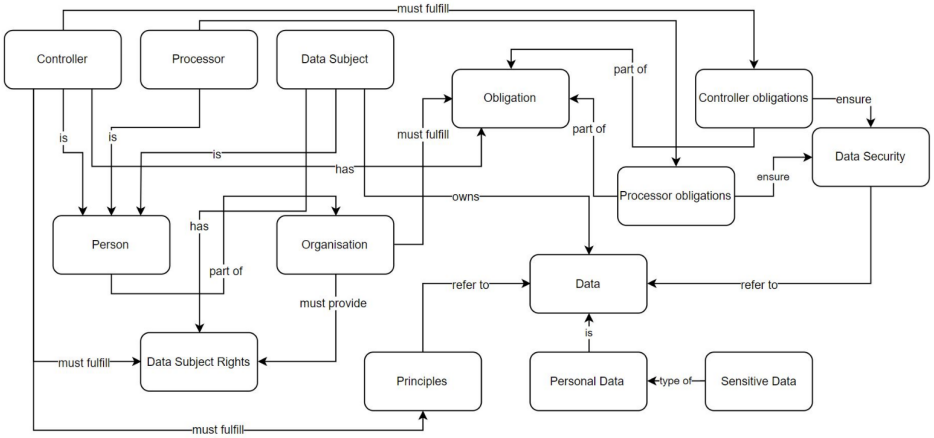


Figure 3.11: Ontology for GDPR and Information Security by Geko & Tjoa [133]

The publication describes an example use-case where additional concepts are added to represent security concepts of access control, audits, data classification, information risk assessment, security awareness, and security policy. Further concepts are also mentioned for representing technical measures of encryption, logging, network security, physical security, privacy by design and default, and pseudo-anonymisation. The publication explicitly mentions that the ontology has not been evaluated. The publication also does not provide access to the ontology or its documentation.

3.3 Other approaches addressing GDPR compliance

Layered Privacy Language (LPL)

LPL [134] presents a formal model for specification of legal requirements regarding consent, data provenance, data retention and other privacy related concepts associated with GDPR. The objective of LPL to provide technical systems with a machine-readable privacy policy and privacy by design. LPL defines the following requirements for a privacy language: (i) differentiation between data-source and data-recipient to enable fine-grained access-control; (ii) modelling of purpose-based privacy policies with modelling of data, retention and anonymisation enabling privacy models; (iii) layering of privacy policies to ensure provenance; and (iv) human-readability.

The LPL structure, depicted in Figure 3.12, consists of *LayeredPrivacyPolicy* (LPP) as the root element representing a privacy policy or a legal contract. The element contains attributes for recording version information and an URI to privacy policy containing its human-readable description. Each LPP can refer to an *UnderlyingPrivacyPolicy* to enable tracking of privacy policies across multiple entities. LPP is associated with data sources and its use in (multiple) purposes. The purpose is further associated with data recipients, data, privacy models, and data retention.

LPL contains the concept of *Entity-Hierarchy* where permissions provided to parent entity can be controlled at granular levels through child entities, such as for different departments within an organisation. Similarly, *Purpose-Hierarchy* represents trees of purposes

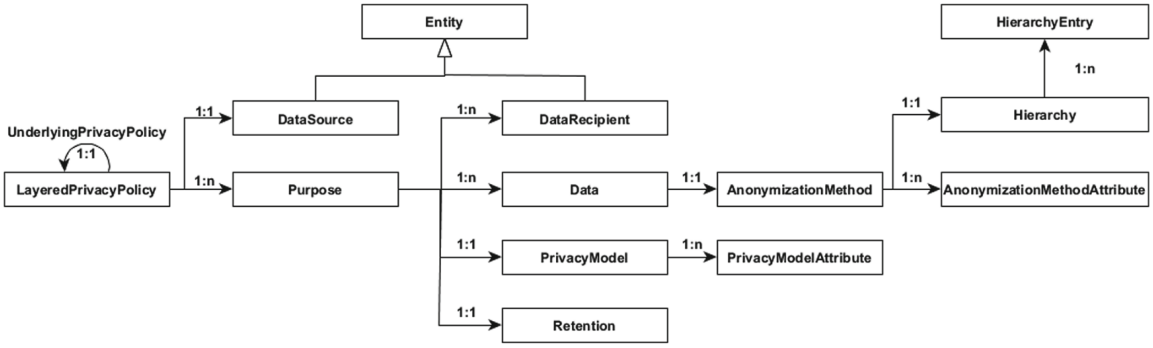


Figure 3.12: GuideMe approach by Ayala-Rivera and Pasquale [134]

where a child purpose inherits rights from parent purpose. In this hierarchy, *Regulated Purpose* specifies purpose required or obligated by law. Authorisation of entities and purposes is carried out by searching for rules matching the entity and purpose or their parents in hierarchy.

A critical analysis of LPL for Article 12-14 of GDPR was carried out [135] and presented as requirements identified in articles of GDPR with indication of whether corresponding LPL concepts fulfil stated requirements. The analysis was used to propose changes to LPL for incorporating additional concepts that fulfil the requirements.

Lodge et al.

Lodge et al. [136] present the Databox application development environment (SDK) for developing IoT apps utilising requirements to meet GDPR compliance. The SDK is provided with the objective to simplify compliance by utilising provided tools and design patterns. The approach is based on identifying applicable requirements from text of GDPR and classifying them as based on risk or transparency.

The SDK enables modelling of applications as information flows consisting of four node types: data-stores - devices or services that generate data; processors - nodes that operate on data; profiler - nodes that infer new information about data subject; and outputs - nodes that perform action on data. Using the SDK, application developers can inspect ongoing risk breakdowns based on developed application, and track personal data as it moves through the application. The SDK also enables creation of GDPR-compliant contracts that embed required information to provide informed consent, and provide automatic tooling for runtime data flow inspection.

Risks are detected based on whether data is exported out of platform, triggers physical actuation, utilises insecure hardware, or uses unverified libraries. Tracking of personal data is provided by enforcing all outputs from a node to provide personal data schema consisting of top-level attributes associated with description of personal data, its source, and sensitivity.

GDPR compliant contracts are created by utilising information flows to construct an agreement that is presented to the user. The agreement is visualised as multi-layered notice that represents information taken from a manifest file containing all possible configurations

in terms of data collection and usage. Developers can mark an information as required or optional to indicate its basis in consent.

Consent and Data Management Model by Peras

Peras [137] presented 10 guidelines for implementing a consent management framework based on requirements of GDPR. These guidelines are based on adopting requirements of consent and provision of rights in a data management system. The publication presents an analysis of existing consent and data management models[26] where elements and attributes of the model are used for comparison. Based on this, a consent and data management model, depicted in Figure 3.13, is proposed consisting of contact interface for interactions between controllers and data subjects, consent management module, context management module, data management module, and origin management.

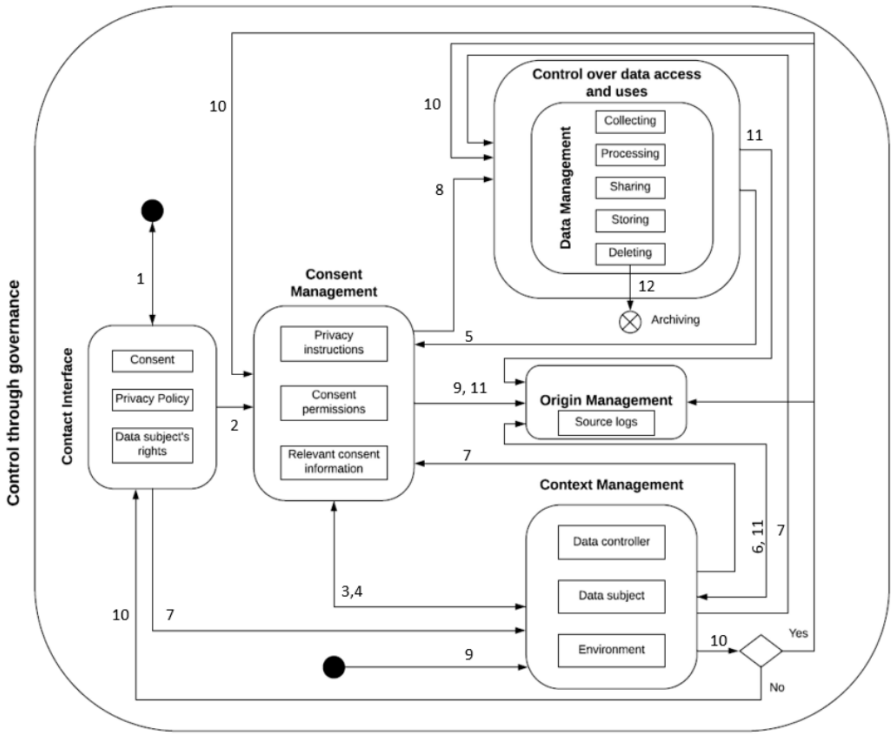


Figure 3.13: Consent and Data Management Model by Peras [137]

Tom et al.

Tom et al. [138] presented a conceptual model of GDPR representing entities, actions, artefacts, rights, and their interactions. The publication provides a table of GDPR articles and their related entities and associations within the model which it uses for evaluating completeness. The model, depicted in Figure 3.14, contains concepts regarding consent, purpose, data processing, technical measures, and actors such as controller, processor, data subject, and supervisory authority.

The attributes of each concept are modelled based on database schemas and represent conditions or additional information associated with the concept. For consent, boolean at-

[26] One of the works cited in this analysis is an approach developed in collaboration by the author [73] which acted as a precursor to the ontology developed to represent consent and presented in Section 5.4.

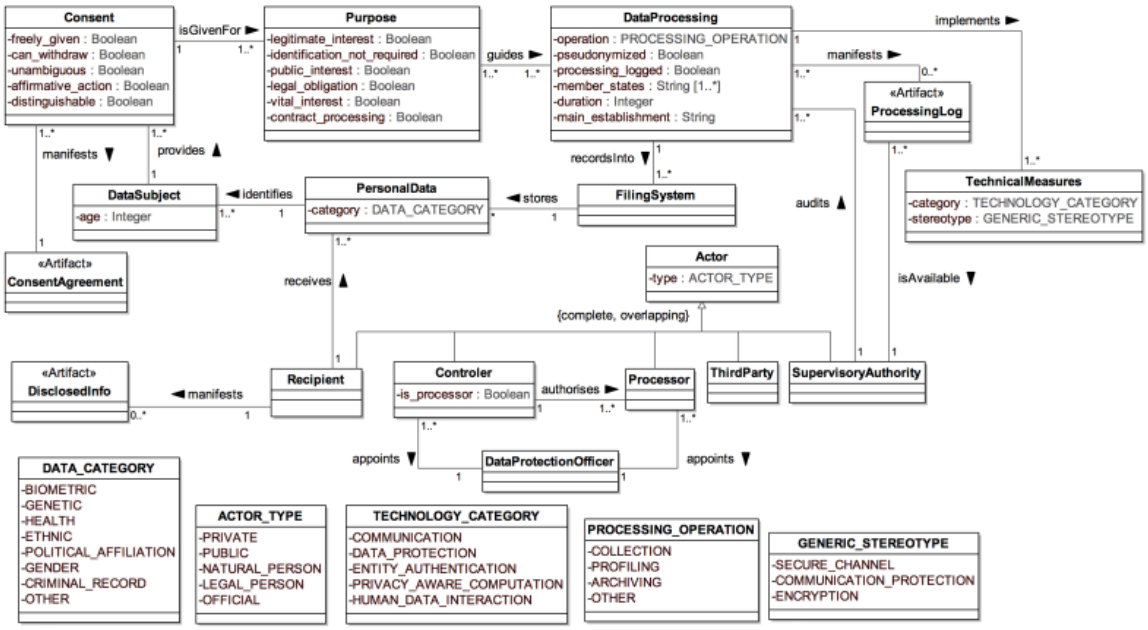


Figure 3.14: Conceptual Model of GDPR by Tom et al. [138]

tributes define whether it was freely given, can be withdrawn, is unambiguous, was in affirmative action, and is distinguishable. Similarly, purpose has boolean attributes regarding whether it is in legitimate interest, does not require identification, is in public interest, for legal obligation, of vital interest to the data subject, and whether it indicates contract of processing. Certain attributes can refer to a table of concepts or instances, such as the *PersonalData* attribute of category referring to *DATA_CATEGORY*. The publication provides examples of personal data categories, actor types (private, public, natural person, legal person, official), technology categories, processing operations, and generic stereotypes (secure channel, communication protection, encryption).

The model was utilised to develop an approach for analysing GDPR compliance over business process models using BPMN [139]. The approach, termed PE-BPMN (Privacy Enhanced BPMN) and depicted in Figure 3.15, utilises BPMN to capture Privacy Enhancing Techniques (PETs) for capturing information flows. The approach is utilised in a mobile app scenario to analyse information disclosure. The information disclosure analysis is performed for identifying whether an object (personal data) is disclosed a defined by it being received or intercepted by another party regardless of intent or policy. Objects can be visible, accessible, or hidden based on whether the actor obtains and owns them, whether data is protected or requires additional permissions, and whether it is protected by some PETs. The associated publication [139] presents details of an implemented prototype and its application in use-cases involving secure storage of data using encryption and PETs.

The approach is similar to the use of semantic web ontologies to represent concepts and use of a semantic reasoner to evaluate compliance, such as in SPECIAL (Section 3.2). The use of BPMN differentiates the approach from other semantic representations of information, and the analysis provides a limited evaluation of compliance based on disclosure of personal data. While the approach contains concepts for representing information associated

with how personal data is processed, it does not provide evidence of its use in evaluating other requirements of compliance - such as analysis of consent, or investigation of activities associated with sensitive personal data.

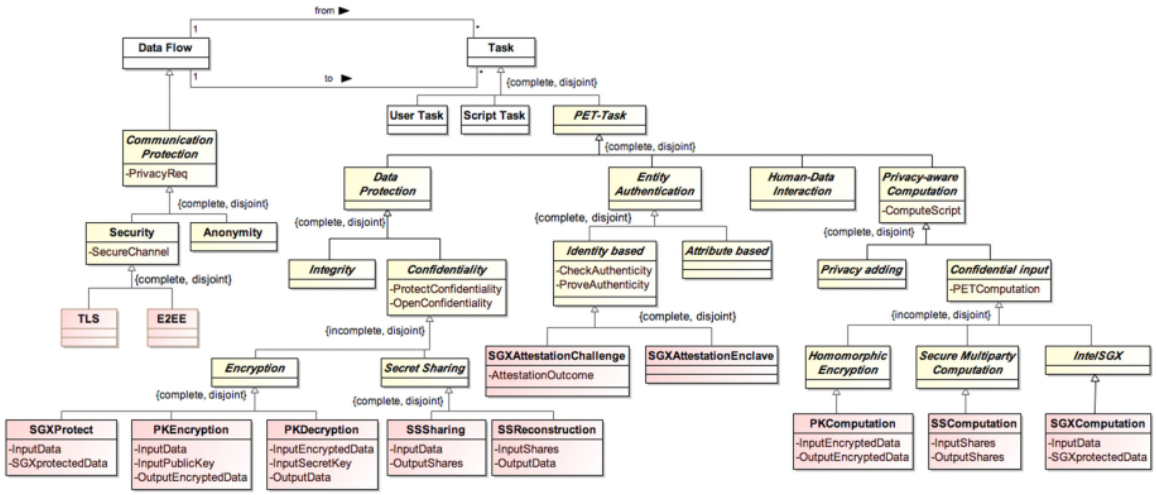


Figure 3.15: PE-BPMN approach by Pullonen et al. [139]

Coletti et al.

Coletti et al. [140] proposed visual design patterns to provide transparency regarding personal data processing in mobile applications. These design patterns consist of recommendations for user interface and experience regarding providing information about processing and rights as per requirements of GDPR. Their approach identifies relevant information to be provided for transparency a: (i) contact information for Controllers, Processors and Data Protection Officers, (ii) information about purpose of use, legal basis, and rights, (iii) information about location as a specific form of data and processing, (iv) information about actions and algorithms utilised in processing, (iv) data sources, and (v) period of use and storage of data.

The visual design patterns organise this information into different layouts and screens to provide a simple view to data subjects. A 'springboard' consisting of overview of different sections is provided through which information about each topic can be obtained. The approach recommends a three-layered approach consisting of - (1) purpose, legal basis, and rights; (2) subgroups of information such as specific laws related to legal basis, actionable tasks regarding rights; and (3) textual information with a search field.

Corrales et al.

Corrales et al. [141] propose embedding legal requirements in smart contracts as a programmatic algorithm by encoding answers to a questionnaire as a compliance framework for data processing in the cloud. The questionnaire contains 15 questions related to legal and technical issues including privacy, data protection, and data security. The answers to questions are boolean and consist of affirmative (YES) or negative (NO). The algorithm for compliance checks whether questions have been answered in a way that meets requirements of GDPR.

The pseudo-code presented in the publication consists of steps checking answers which can also be visualised as a flow-chart or decision making steps.

The approach is based on assessment of questionnaire responses rather than analyses of process representations. While questions used are useful in investigation of compliance, the approach does not provide any guideline or assistance regarding gathering information required to answer them, or even evaluating information apart from boolean responses.

LUCE

LUCE [142] (A Blockchain Solution for monitoring data License accoUntability and CompliancE) is a platform for facilitating compliance with licensing terms by enabling data accountability through recording of data use and purpose using blockchain. The platform specifically aims to provide compliance assistance for GDPR and its rights. LUCE has objectives to : (i) automatically manage and enforce licensing terms attached to datasets, ii) record and make available information pertaining to how and for which purpose datasets are reused, and iii) enable compliance with GDPR rights regarding data access, rectification, and erasure. The architecture, depicted in Figure 3.16, consists of four layers for management of metadata, entities, data, and blockchain.

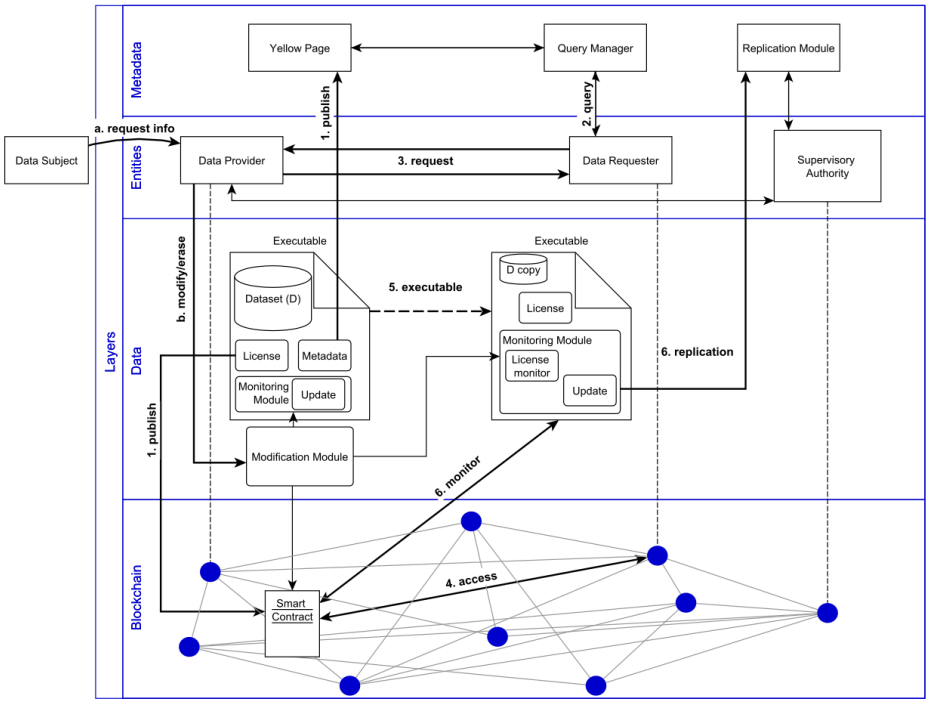


Figure 3.16: Architecture of LUCE [142]

Data providers publish their dataset metadata into a *Yellow Page* shared directory where requesters can search for matching datasets. Each dataset is accompanied with licensing information for access and reuse. A transaction for reusing data is implemented as a smart contract containing associated licenses as obligations. Requesters accepting licensing agreements are provided with access tokens which are tracked within smart contracts for compliance. The *License-Monitoring* sub-module continuously checks actions performed by data requester for compliance with licensing terms and records actions in a file to record log of events. Access tokens are valid for specific time periods after which renewal is granted

based on prior compliance with licensing terms checked using event logs. A data provider can also check accesses to data and status of requester's compliance. GDPR rights are enforced and supplemented by using smart contracts to perform actions such as retrieving information to provide regarding right to access and obtaining IDs for deletion regarding right to erasure and rectification. The prototype implementation described uses Ethereum[27] as blockchain platform and Solidity[28] as smart contracts language.

Decision Provenance by Singh et al.

Singh et al. propose the concept of 'Decision Provenance' which uses methods based on provision to identify decision pipelines and actions taken in an system at design and run-time. Decision provenance is aimed to facilitate oversight, audit, compliance, risk mitigation, and user empowerment, and relates to accountability of algorithmic systems. The paper provides a discussion of GDPR, identifying legal impetus to identify and record provenance information regarding processing of personal data and management of rights. The utilisation of data flows in a system to map actions, steps, and boundaries is utilised as motivation for its use in compliance management systems based on adapting legal requirements for design and use of data flows. With this, 'decision provenance' is defined as "use and means for provenance mechanisms to assist accountability considerations in algorithmic systems".

The information required to be recorded and utilised for decision provenance concerns the history or provenance of data and a broader view of system behaviour and interactions including those with entities. The publication also describes usefulness of information in ex-post analysis. Regarding compliance, decision provenance is described with uses for tracking of conditions associated with processing, consent, purposes, rights, and sharing. The collected information is also described for uses in performing audits and investigations by authorities as well as the organisation itself.

The publication makes specific mention of PROV-O [47] for representing provenance information, supplemented by ontologies for describing additional information about data processing. An example cited in this context concerns the use of PROV-O to represent provenance for GDPR compliance (Ujcich et al. [50] - see Figure 3.2).

Sion et al.

Sion et al. [143] presented an approach for data protection by design in software architectures towards incorporating requirements of GDPR. The approach is developed towards automation of data protection impact assessment (DPIA) using (i) meta-model constraints, (ii) model analysis, and (iii) interaction with stakeholders. It uses architectural notations such as UML and Data Flow Diagrams (DFD) to express the model of a system. The publication provides a discussion of necessary information for creation of a meta-model for GDPR which includes the following:

- Use of GDPR terminology as opposed to abstract architectural notations;
- Core abstractions such as *ProcessingPurpose*;
- Documentation regarding abstractions - such as collection, further processing, the purpose itself, its legal basis (termed *LawfulGround*), involved actors and their roles, and

[27] https://ethereum.org/
[28] https://solidity.readthedocs.io/

datasets and data-types involved for a purpose.

The presented meta-model, depicted in Figure 3.17, utilises UML to denote concepts and relationships. Apart from expressing concepts such as purpose, processing, personal data type, and data subject type, the meta-model expresses controllers, processors, recipients and third parties as legal roles expressed through an actor. Additionally, it has a distinct concept for further processing to represent processing other than that of primary purpose. Exceptions to obligations provided by GDPR are captured as prohibition and exemption types.

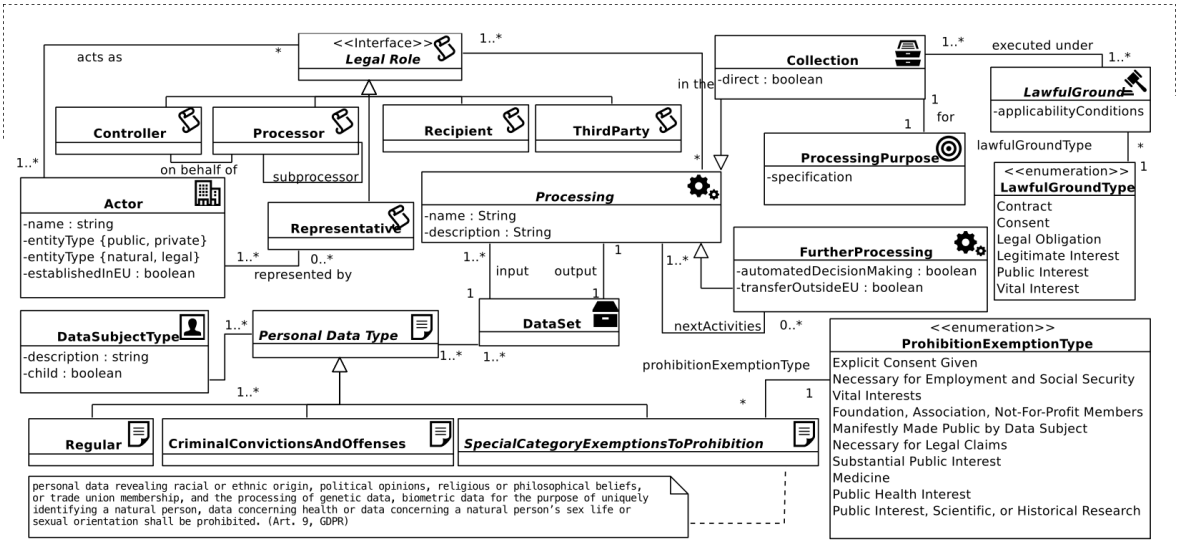


Figure 3.17: Meta-model of architectural view for GDPR by Sion et al. [143]

The meta-model is used to express the model or architecture of a system which is then evaluated or validated utilising constraints. One of these, described in the publication, is "soundness criteria" - which is relevant to completeness and correctness of a model. Examples of these are - "every chain of Processing needs to start with a Collection", and "Every input DataSet to a Processing needs to be the output DataSet of a Processing that is located earlier in the chain of Processing". These can be collectively expressed as constraints for every data used in processing to have a source whether by collection or as output of processing activity. The publication also provides examples of incorporating legal requirements in the meta-model, with examples for purpose limitation, minimisation, and automated decision making involving special categories of personal data.

The approach is validated through an use-case in health domain involving a patient monitoring system (PMS) using DFD to denote architectural notations in Eclipse Modelling Framework with graphical visualisations created using Sirius viewpoint specification. The soundness checks were performed using Acceleo Query Language supported by Sirius, with further described as implementing compliance assessment rules using VIATRA's graph-based pattern language.

privacyTracker

privacyTracker [144] is a framework that provides data traceability through privacy by design principles for GDPR compliance. The framework utilises policies to mediate access

to information stored in as *Customer Records* and consists of 3 modules regarding data collection, data distribution, and data traceability. The *Customer Record* is a XML document utilising XSD to store personal data and contains two sections - a mandatory metadata section and an optional section. The metadata section contains fields for record identification, data tractability, and cryptography controls. The optional section consists of fields indicating public data, private data that can be disclosed based on consent, and data provided by the enterprise itself.

The record identification field contains a URI based on string concatenation of company name, user email address and auto-generated random identifier. This URI is unique within the framework but changes when data is distributed to another entity. In addition, the identification field also stores timestamps associated with record genesis, (local) record creation, and expiration. The data tractability fields record links to the entity that generated the record (backward-to-root reference), entities the record was obtained from (backward reference), and entities the record is disclosed to (forward reference). The cryptography controls store a signed copy of received record and a signature as hash code of complete record. The distribution module enables sharing of data via an API through which granular requests can be made. Records of distributions are signed cryptographically and verified for tractability. Enforcing rights is simplified by traversing the stored distribution records starting from original record as root and moving towards leaves. A prototype of implementation consisted of 6 companies and used MySQL and PHP as technological framework. The records were stored using XML are then ingested into the database and queried using SQL.

The analysis of GDPR presented in connection with privacyTracker provides a list of requirements required to be satisfied by compliance frameworks, which are: Articles 5(1a), 5(1d), 6(1a), 6(1c), 7(1), 7(3), 12(1), 12(2), 14(1a), 14(1ac), 14a(2g), 15, 16(1), 17(1), 17(2a), 17a(1), 18(2), 19(2)). These provide obligations to record and demonstrate evidence regarding handling and sharing of consumer data.

Metrics for Transparency by Spagnuelo et al.

Spagnuelo et al. [145] define eight quality metrics for transparency regarding data processing associated with GDPR. In this context, transparency is defined based on quality factors of 'Informativeness', 'Understandability', and 'Accessibility' associated with providing information; and 'Validity' and 'Accessibility' associated with providing mechanisms. Each metric is associated with one or more questions that retrieve information regarding quality factors associated with quantitative scoring. The metrics provide a quantifiable representation of transparency in a system and are used in a use-case to evaluate Microsoft HealthVault - a commercial product. The metrics are as follows - accuracy, currentness, conciseness, detailing, readability, availability, portability, and effectiveness.

meta-model for PLA by Diamantopoulou et al.

Diamantopoulou et al. [146] present a meta-model for representing Privacy Level Agreements (PLA) with the aim of establishing contracts between controllers and individuals. The meta-model, depicted in Figure 3.18, represents privacy preferences as questionnaires where each question refers to specific data categories and answers represent individual's

preferences regarding data sharing. The meta-model associates preferences with threats, risks, and security measures through data and controller. Requests for sharing data are associated with data and preferences, and can be accepted or denied based on values of linked preferences. The meta-model is specified to be designed for use-case in eGovernance scenarios where public administrators can provide data and value based on citizens preferences represented through the meta-model.

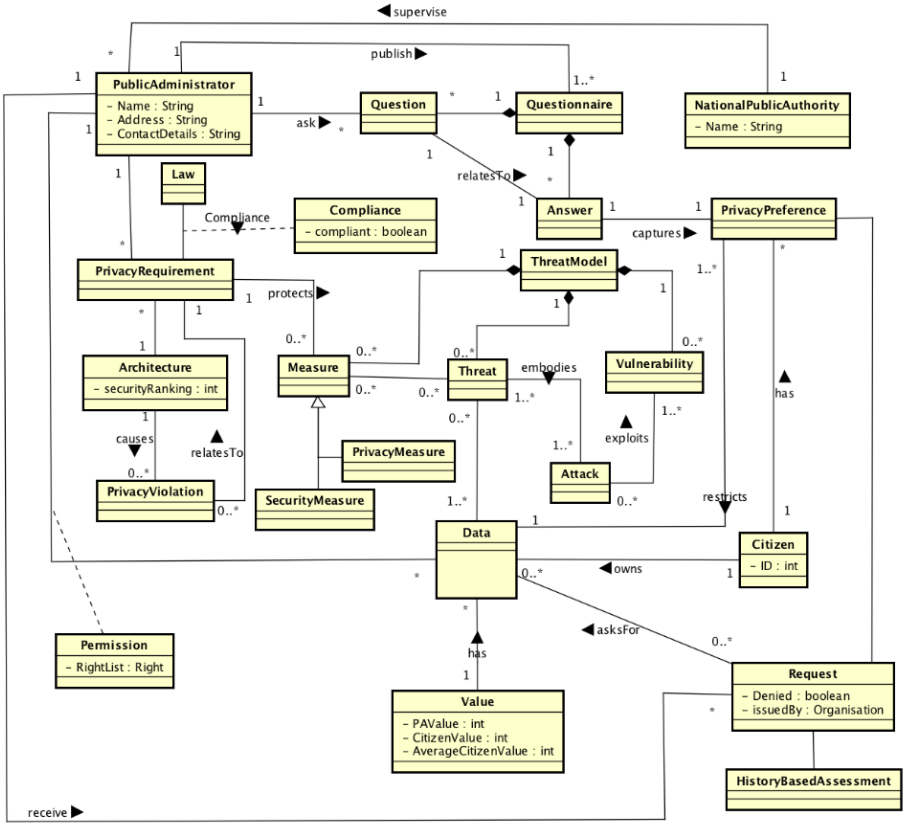


Figure 3.18: meta-model for Privacy Level Agreements by Diamantopoulou et al. [146]

Robol et al.

Robol et al. [147] proposed a modelling and reasoning framework based on socio-technical systems [148] - an approach for incorporating interaction between people and technology. The proposed framework extends STS-ml - a goal-based modelling language provided in STS - with relevant concepts regarding privacy by design and GDPR compliance. The extended modelling language consists of three views - social, information, and authorisation - based on concepts and context of interactions between them. Social views incorporate actors and their goals and documents, as well as delegations and transmissions. Information views consist of associating actors with information and tangible documents. Authorisation views consist of interactions with actors based on authorisations which are validated by legal basis over some information towards a particular goal. The modelling concepts are explained using a health-domain use-case.

Reasoning is based on using concepts in policies that are validated for compliance. The policies use formal representations of rules based on conditions and constraints for compliance with the notion of *well-formedness* of models. One notable aspect of this work is use of

concepts from socio-technical approaches to represent GDPR concepts. In particular, goal is used to simulate purpose, actors are used to conceptualise individuals and controllers, and authorisation is used to simulate valid legal basis. The approach enables use of STS modelling and reasoning to develop rules for expressing GDPR compliance requirements and validations.

Basin et al.

Basin et al. [149] proposed an approach that aligns purpose as defined by GDPR with business processes and uses formal models of inter-process communication to audit GDPR compliance. The approach also uses models of data flows to derive privacy policies. The definition of business process used consists of having a set of activities that take inputs and produce outputs, have a beginning and an end, and are ordered. The inter-process communication is based on data-flow graphs where different processes interact and share same sets of data for different purposes.

The auditing mechanism is based on validating whether an implementation conforms to process collections i.e. implemented activities conform to modelled processes, process collection conforms to privacy policy i.e. modelled processes conform to those specified within a privacy policy, process collection conforms to GDPR, and privacy policy conforms to GDPR. The process collection is defined as a tuple $PC = (P, D, DU, DC)$ consisting of processes (P), data classes (D), data usage ($DU \subset DxP$), and data collection ($DC \subset DxP$). Consent compliance is based on evaluating whether process collection permitted by consent is permissive based on the defined privacy policy. Similarly, data minimisation is evaluated by assessing whether all data collected is utilised in some process.

RestAssured

RestAssured[29] is an European H2020 project that aims to enable secure data processing in the cloud with sticky policies for decentralised data life-cycle management. The project aims to achieve this by using run-time models for data protection assurance and automated risk management. Run-time detection, prediction, and prevention of data protection violations is achieved using a catalogue of risk patterns representing data protection vulnerabilities of cloud systems [150], [151]. The pattern meta-model and its validation is described in deliverable D5.1 [152]. Information about the project is available through peer-reviewed publications and deliverables[30].

The project addresses GDPR through a conceptual model incorporating concepts necessary for compliance as depicted in Figure 3.19 [153]. The model is utilised in data protection policy specification consisting of data protection contract, sticky policies, and specification of obligations. The model is utilised to provide a customised privacy policy that allows users choice of preferences for selected services [154]. The data protection contract is composed of 4 main parts - description of service that needs to access data, specification of data needed to be accessed by service and whether access is mandatory, list of usages where usage is composed of textual description that describes purpose of usage and associated processing action as well as list of needed data types, and specification of data that will be published by

[29] https://restassuredh2020.eu/
[30] https://restassuredh2020.eu/publications/

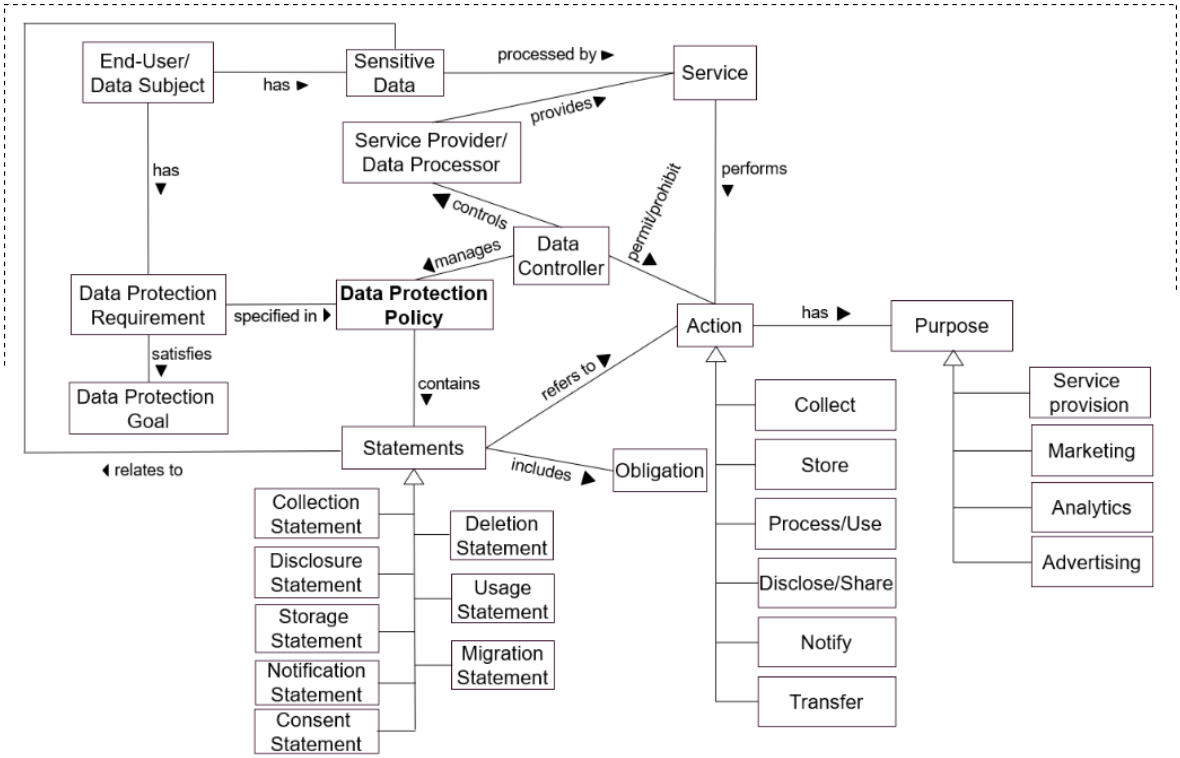


Figure 3.19: Model of policy specification framework in RestAssured [153]

service. The obligations are defined based on their enforcement in relation to access control (before, at same time, after), and are specified with three pieces of information: (i) obligation type (before, after, with), (ii) event trigger, and (iii) action to be performed. The serialisation of obligations is done using XACML 3.0. The deliverable D7.1 [155] refers to use of RDF to represent concepts and relationships as a graph consisting of three layers representing: (i) core model of schema that captures fundamental concepts e.g. assets, roles, threats, and their relationships; (ii) domain model of datasets that encode domain-specific knowledge e.g. detailed threats and their possible control strategies; and (iii) system models that represent an actual system upon which risk assessment is being performed.

OPERANDO

OPERANDO[31] (Online Privacy Enforcement, Rights Assurance & Optimisation) is an European H2020 project that aims to provide a platform for users to specify preferences via a dashboard which is then compared with online service providers' privacy policies and converted into access control rules. The deliverable D3.1 [156] describes legal requirements of GDPR in context of OPERANDO's objectives, while deliverable D6.4 [157] describes architecture and working of privacy tools. Information about the project is available through peer-reviewed publications and public deliverables[32].

The platform operates in an online environment and uses APIs between services which pass information using JSON. A User Privacy Policy (UPP) [158] is stored in platform database and represents user's preferences (and consent). The UPP records preferences using fields

[31] www.operando.eu/
[32] http://www.operando.eu/servizi/moduli/moduli_fase01b.aspx

for information identifiers, personal data category, ranked preference (0 to 10, higher indicating greater concern), role (of person acting on data), action (performed on data), purpose, and recipient. The UPP also contains information on access policies created based on comparing user's preferences against an online service provider's privacy policy.

The online service provider's privacy policy (OSP) [158] consists of fields defining workflows and contained steps with information about requester subject (role of service provider), requested data, and action to be performed. The OSP also contains information about access policies for which service provider has privileges to perform requested operations in specified roles. Information about how UPP and OSP are utilised in an API to enact an access request is provided through deliverable D6.4 [157].

My Health My Data (MHMD)

My Health My Data[33] (MHMD) is an European H2020 project that aims to develop infrastructure based on blockchain and smart contracts to provide personal data accounts in the cloud that can be managed using dynamic consent interfaces and to provide peer-to-peer connections between stakeholders. The project uses blockchain to log data transactions in a secure, transparent, and accountable manner, with de-identification and encryption to protect identity and sensitive information. The safety and security of data is tested using re-identification and penetration simulations. Information about the project is available through peer-reviewed publications and public deliverables[34].

The project is applied over use-cases of health data and devices as described in deliverable D1.1 [159], which uses an ontological resource to model the common data ontology - described in deliverable D4.2 [160]. The common data ontology consists of modular health data ontologies representing synthetic data shared by the project's commercial partners as an use-case. Access to data is provided through an API that includes parameters describing requested data category and consent. The API returns matching datasets which can then be utilised for data processing activities. The consent description parameter of API is a text field consisting of values such as "synthetic data" and "fully anonymised" which reflect state of data and its requirements in terms of consent.

Though the project aims to utilise GDPR as a source of legal requirements and strives to design its framework to meet compliance requirements by both design and default, there is no publicly available information regarding specifics of how GDPR compliance is achieved or represented. It focuses on privacy preserving and security aspects of data storage by using technological solutions such as access control and transparent logging to design specifications based on legal requirements. The project does explores impact of GDPR on storing data in a blockchain, especially regarding right to be forgotten [161].

3.4 Approaches involving Privacy Policies

These approaches involve utilisation of privacy policies for various purposes associated with data protection and user empowerment. While they do not target GDPR compliance, they are relevant as they involve analysis of information associated with compliance - such

[33] http://www.myhealthmydata.eu/
[34] http://www.myhealthmydata.eu/publicdeliverables/

as purposes, data categories, and rights - and provide an overview of GDPR's impact and uptake in the real-world.

Usable Privacy Project

The UsablePrivacy project [162] utilised natural language programming over manually annotated corpus of privacy policies to create a system for automatic classification of policies. It utilised semantic web technologies [163] to represent annotations and information within a privacy policy, and to query information for retrieval. Its datasets have been utilised in other similar approaches [164], [165] concerning automatic analyses over privacy policies.

Polisis and other approaches based on Usable Privacy Project

Galle et al. [166] presented an argument for a similar corpus (to UsablePrivacy) of GDPR-specific privacy policies based on increased requirements of GDPR compliance, presence of information, and use of simpler text. While such a dataset has not been published to date, there have been approaches incorporating GDPR specific information in privacy policy analysis. Polisis [164] is one such approach that analyses privacy policies using Wilson et al.'s taxonomy of privacy concepts [167] (depicted in Figure 3.20). Further GDPR specific work using the same approach collects and compares policies in pre-GDPR and post-GDPR versions to identify impact of GDPR [165]. The approach uses a list of queries derived from ICO's GDPR compliance checklist to obtain coverage of GDPR in terms of queries answerable in privacy policies.

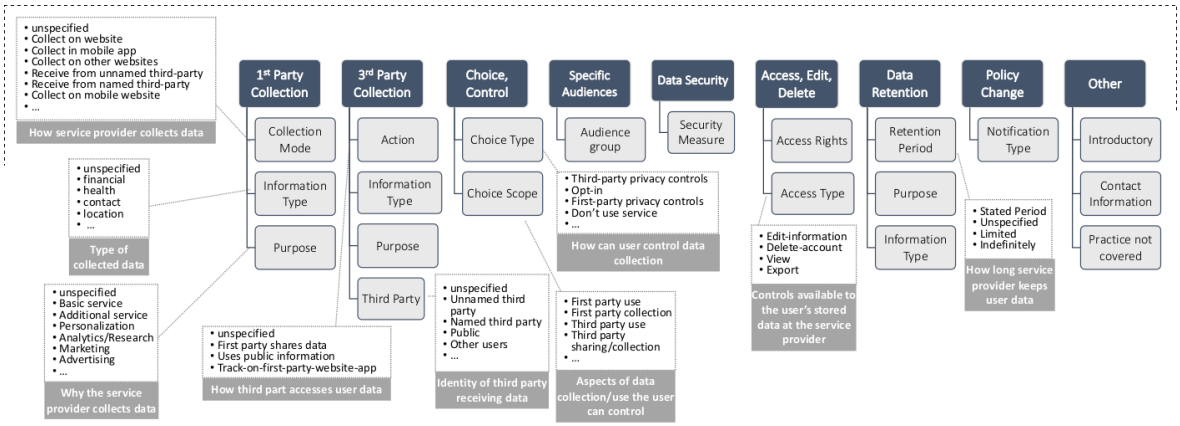


Figure 3.20: Privacy taxonomy by Wilson et al. [167] used by Harkous et al. [164] and Linden et al. [165] to analyse privacy policies

PrivacyGuide

PrivacyGuide [92], [168] provides visualisation of privacy policies and information categories as a dashboard. It uses a similar approach to UsablePrivacy and Polisis in classifying statements in policies and extracting information using machine learning. In terms of information, its models contain concepts for: data collection, protection of children, third-party sharing, data security, data retention, data aggregation, control of data, privacy settings,

account deletion, privacy breach notification, policy changes. The approach is based on identifying a set of keywords for each concept and classifying statements based on their presence.

3.5 Approaches related to Consent

These approaches concern representation of consent and provide information about existing efforts towards its management and analysis. They are relevant as they involve information associated with provision of consent, its impact in real-life, and complexities associated with its representations.

Grando et al.

Grando et al. [169] present an ontology-based model of permissions and obligations resulting from informed consent within medical domain. The approach, depicted in Figure 3.21, consists of using XACML for rules implementing a consent management system by utilising reasoning using SWRL.

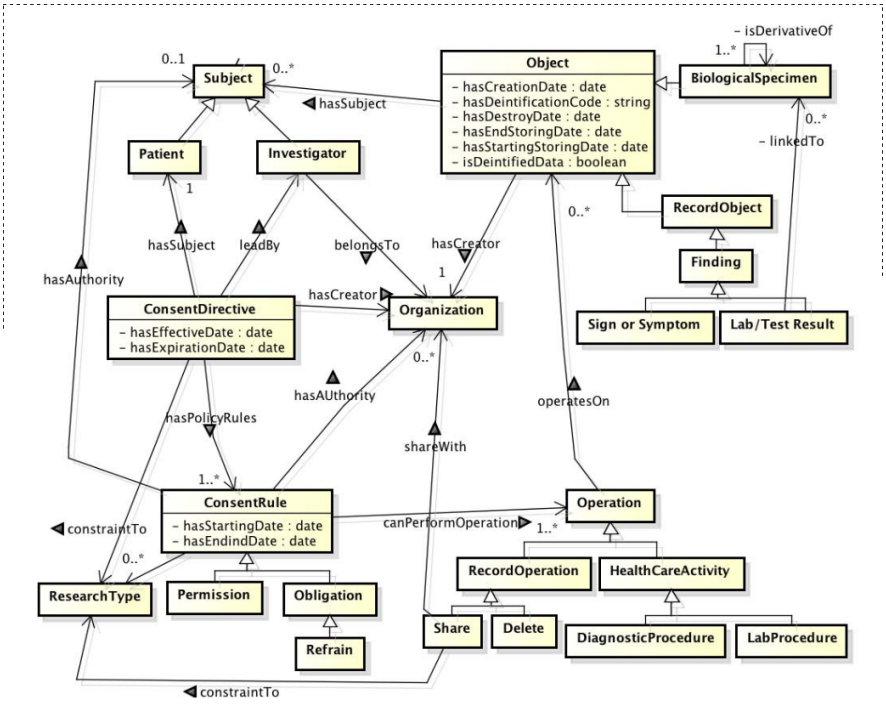


Figure 3.21: Informed consent permissions ontology by Grando et al. [169]

Consent Receipt Specification

The Consent Receipt specification[35] [170] is a standardisation effort by Kantara Initiative - a non-profit industry professional trade association with a mission to - "improving trustworthy use of identity and personal data through innovation, standardisation and good practice in the domain of digital identity management and data privacy". Consent receipt provides a

[35] https://kantarainitiative.org/confluence/display/infosharing/Consent+Receipt+Specification

list of fields for capturing context of given consent and is meant to provide a record of transaction regarding personal data. Its fields, depicted in Figure 3.22, capture information about timestamp the consent was given at, entities involved (controllers), PII involved, purposes, and recipients.

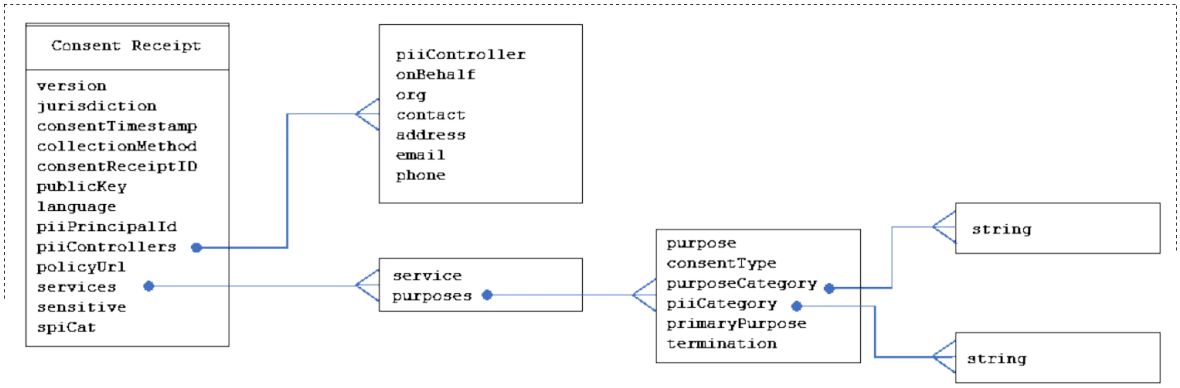


Figure 3.22: Fields in Consent Receipt [170]

3.6 Upcoming Research Projects Addressing GDPR

This section lists upcoming research projects that address or incorporate GDPR compliance in their requirements and are currently developing solutions to meet their objectives and aims. These are presented to include them in the larger context of approaches associated with GDPR compliance, but are not analysed as part of SotA relevant to this thesis given their nascent research output.

PDP4E

PDP43[36] (Methods and Tools for GDPR Compliance through Privacy and Data Protection Engineering) is an European H2020 project that aims to provide tools and guidance for incorporating privacy and data protection into software development life-cycles (SDLC) to implement data protection by design. Currently, the project lists one deliverable on its website, D2.1 [171], which provides information on legal analysis of GDPR in terms of requirements and obligations for software, and identifies roles and responsibilities within for SDLC.

DEFeND

DEFeND[37] (Data Privacy Governance for Supporting GDPR) is an European H2020 project that aims to create a platform for providing services to organisations regarding GDPR compliance. The platform will provide services regarding data management and governance, data breach reporting, process management, and GDPR planning and reporting. The peer reviewed publication [172] presents more details about the architecture and its components.

The Data Assessment Component (DAC) consists of Organisation Data Collection (ODC) module which uses an questionnaire to collect information about organisational scope, data

[36] https://www.pdp4e-project.eu/
[37] https://www.defendproject.eu/

processing, processes, and activities - and is used to evaluate the organisation with relevant aspects of GDPR. The questionnaire responses from ODC are given to Assessment Translator (ATr) module which converts them to an XML-schema in order to create Data Assessment Model (DAM) - a goal-based requirement engineering model of actors, assets, establishments and data flows.

The DAM is used by Data Privacy Analysis Component (DPAC) to perform analysis about DPIA, data minimisation, privacy-by-design/default, and threat mitigation. The outcome of analysis is a Data Privacy Model (DPC) which is utilised to specify and evaluate at design and run-time the operations of consent management, data access rights, security and privacy technologies. The project also aims to create a dashboard to provide organisations with control and monitoring of operations to achieve GDPR compliance, and to enable data subjects to interact with the platform based on consent.

PAPAYA

PAPAYA[38] (PlAtform for PrivAcY preserving data Analytics) is an European H2020 project that aims to provide privacy preserving techniques and technologies for performing analytics tasks on encrypted data by untrusted third-party data processors. The project uses obligations of GDPR to structure outcomes in terms of Privacy Enhancing Technologies (PETs) which are provided through an interoperable platform. The project focuses on application of privacy by design principle to provide usability, transparency, and auditability to end users regarding processing of their data. Information about the project is available through peer-reviewed publications[39].

SMOOTH

SMOOTH[40] is an European H2020 project that aims to assist micro enterprises become compliant with GDPR by designing and implementing tools for awareness about GDPR obligations and analysing their level of compliance. The project objectives include creation of a platform for automatic assessment of privacy protection documents (including privacy policies), stored data, and processing of personal data on websites or mobile apps to create a compliance report. Further details about the project are currently not available.

SODA

SODA[41] (Scalable Oblivious Data Analytics) is an European H2020 project that aims to provide tools for performing analytics over encrypted data at large scales while providing compliance with GDPR. The project will utilise use-cases within health domain to demonstrate privacy-preserving aspects of its technologies. SODA incorporates requirements of GDPR and provides guarantees regarding compliance through use of its tools and technologies. The analysis of legal requirements arising from GDPR and role of SODA in addressing them

[38] https://www.papaya-project.eu/

[39] The deliverables listed at https://www.papaya-project.eu/deliverables are not publicly accessible at the time of writing this thesis.

[40] https://smoothplatform.eu/

[41] https://www.soda-project.eu/

is presented in deliverable D3.1 [173]. Further information about the project is available through peer-reviewed publications and public deliverables[42].

DECODE

DECODE[43] is an European H2020 project that aims to provide tools for data ownership by using blockchain with attribute-based cryptography. The project will be piloted in Amsterdam and Barcelona and will focus on enabling citizens to manage and control data generated through IoT devices. The project objective is to provide data as a shared resource for collective benefit and crowd-sourced information. Smart contracts will be used to enable control of data and will be stored in a distributed ledger (such as blockchain). Information about the project is available through its public deliverables[44].

The specification of information required for legal compliance as well as processing is defined in terms of 'entitlements', which include personal data category, description, purpose, condition (alternate purpose), expiry date (storage duration). These are collected across use-cases, normalised to find commonalities, and utilised in smart policies to record information about processing. The deliverable D3.5 [174] describes collected entitlements from pilot use-cases in Barcelona. The analysis of legal requirements and their incorporation in the DECODE project is described in D1.8 [175].

MOSAICrOWN

MOSAICrOWN[45] (Multi-Owner data Sharing for Analytics and Integration respecting Confidentiality and OWNer control) is an European H2020 project that aims to enable data sharing and collaborative analytics in multi-owner scenarios. Its objectives are to provide a data governance framework that captures and combines protection requirements specified by multiple parties, and provides protection techniques for enabling efficient and scalable data sharing and processing. The project intends to create an approach for combining selective encryption, blockchain, and smart contracts to enable data owners to leverage data markets to monetise their data in a con-trolled way.

PoSEID-on

POSEID-on[46] (Protection and control of Secured Information by means of a privacy Enhanced Dashboard) is an European H2020 project that aims to safeguard rights of data subjects while simultaneously supporting organisations in data management and processing by ensuring GDPR compliance. Its objective is to create a privacy enhancing dashboard for personal data protection using an implementation of permissioned blockchain and smart contracts to provide accountability, transparency, and compliance. The project aims to develop technologies for automated detection of unexpected and potentially harmful behaviours in order to monitor privacy risks and notify threats to data subjects during transactions. Information about the project is available through its public deliverables[47].

[42] https://www.soda-project.eu/deliverables/
[43] https://decodeproject.eu/
[44] https://decodeproject.eu/publications
[45] https://mosaicrown.eu/
[46] https://www.poseidon-h2020.eu/
[47] https://www.poseidon-h2020.eu/documents/

Deliverable D2.1 [176] provides information on use-cases, user stories, personal data required, and third parties involved. Deliverable D3.1 [177] presents the architecture with information about smart contracts utilised on permissioned blockchain to support permissions regarding requesting, granting, revoking, notifying, checking, and accessing permissions regarding use of PII. Each permission event is logged to blockchain and consists of data processor, data subject, and personal data field involved. Deliverable D4.3 [178] describes use of natural language programming approaches to detect PII in stored data as a measure of risk detection and management.

3.7 ANALYSIS

3.7.1 Overview of SotA

The analysis of state of the art was carried out using the methodology presented earlier in Section 3.1. The approaches were first analysed to identify relevance for categories of information identified in the methodology. Table 3.1 presents this information by using a check mark (✓) to indicate the approach uses or provides information about that topic, with its absence indicating that the approach does not address that topic or no information about its use could be found. The column headings represent identified categories of information, and represent - (T): type of work where PRJ indicates a research project and RES indicates research publication, (R): if it models GDPR clauses and concepts, (M): if it provides an ontology, (EA): if it represents information for ex-ante Compliance, (EP): if it represents information for ex-post compliance, (P): if it models activities, (C): if it represents consent, (Cm): if it evaluates compliance, (Rq): if it specifies requirements for GDPR compliance, and (O): it provides resources in an open and accessible manner.

Of the 44 approaches listed in the table, upcoming research projects are not considered part of state of the art and are therefore not analysed. The 5 approaches related to privacy policies and 2 approaches related to consent do not directly address GDPR compliance, but provide representation of information relevant to modelling of information. The analysis is primarily carried out over 29 approaches which directly address GDPR compliance. The following observations are made based on an overview of Table 3.1:

- **Research Projects**: Of 29 approaches presented as addressing GDPR, 9 are outcomes or part of a research project (column *T*). Of these 9 projects, 6 approaches are part of 12 that utilise semantic web and feature cross-collaboration in their research and outputs. If cross-collaborations between projects are consolidated under the project providing primary resources for GDPR compliance, 2 projects - SPECIAL and MIREL - are notable in their scope of utilising semantic web and produced resources.

  The research projects are primarily funded by European Union research grants for addressing or incorporating GDPR as a requirement. Given that the GDPR is an European regulation, it is not surprising to have research projects funded by EU for researching applications of technology for compliance. These projects represent dedicated resources in terms of funding and personnel, and feature collaboration between universities, research institutes, and industry - which includes participation of legal experts, authorities, and legal domain organisations.

  In terms of output, deliverables of research projects represent a wider consensus be-

Table 3.1: Overview of approaches in SotA

Work	T	R	M	EA	EP	P	C	Cm	Rq	O
Approaches using semantic web to address GDPR compliance										
SPECIAL	PRJ		✓	✓	✓	✓	✓	✓		✓
SERAMIS	PRJ	✓	✓					✓	✓	
Vos et al	RES		✓					✓	✓	✓
CitySPIN	PRJ		✓	✓	✓	✓	✓	✓		✓
MIREL	PRJ	✓	✓	✓	✓	✓		✓	✓	
DAPRECO	PRJ	✓	✓	✓	✓	✓		✓	✓	
BPR4GDPR	PRJ			✓	✓	✓		✓	✓	
Elluri et al.	RES		✓							✓
Ujcich et al.	RES		✓		✓	✓				
PICS	RES				✓					
AdvoCATE	RES		✓				✓	✓		
Geko & Tjoa	RES		✓							
Other approaches addressing GDPR compliance										
LPL	RES		✓							
Lodge et al	RES		✓		✓	✓	✓	✓		
Peras	RES		✓				✓			
Tom et al	RES		✓	✓		✓		✓		
Coletti et al	RES		✓	✓			✓		✓	
Corrales et al	RES							✓	✓	
LUCE	RES					✓		✓		
Singh et al	RES			✓						
Sion et al	RES		✓	✓		✓		✓		
privacyTracker	RES		✓		✓	✓		✓	✓	
Spagnuelo et al	RES								✓	
PLA meta-model	RES		✓							
Robol et al	RES		✓	✓				✓		
Basin et al	RES			✓		✓		✓		
RestAssured	PRJ		✓	✓	✓	✓	✓	✓		
OPERANDO	PRJ						✓	✓		
MHMD	PRJ						✓			
Approaches involving Privacy Policies										
UsablePrivacy	PRJ		✓							✓
Galle et al	RES		✓							
Polisis	RES		✓							
Linden et al	RES		✓							
PrivacyGuide	RES		✓							
Approaches related to Consent										
Grando et al.	RES						✓			
Consent Receipt	PRJ						✓			✓
Upcoming research projects addressing GDPR										
PDP4E	PRJ								✓	
DEFeND	PRJ		✓	✓				✓		
PAPAYA	PRJ									
SMOOTH	PRJ									
SODA	PRJ							✓		
DECODE	PRJ						✓	✓		
MOSACrOWN	PRJ						✓	✓		
PoSEID-on	PRJ						✓	✓		

tween stakeholders and evaluation in terms of use-cases that are provided by commercial partners. These provide valuable insights into stakeholder requirements and developed solutions within state of the art.

- **Representation of GDPR clauses**: Of 29 approaches addressing GDPR compliance, only three approaches provide a representation of GDPR (column *R*) and do so in a machine-readable format. Of these 3, MIREL and DAPRECO use PrOnto ontology and aligned resources produced by MIREL - and therefore can be consolidated into a single approach using PrOnto.
- **Modelling of GDPR concepts**: 20 of 29 approaches addressing GDPR compliance also model concepts from GDPR as an ontology or data model (column *M*). Additionally, all of 5 approaches involving privacy policies also provide concepts relevant to GDPR. The extent of concepts defined varies between each approach based on its aims and objectives. However, none of these approaches provide a glossary of terms or concepts associated with GDPR compliance. No approach indicates source of its concepts within GDPR except as textual annotations to a clause (e.g. Article 4-11) in some cases.
- **Representing activities**: 13 approaches represent and utilise information about activities associated with GDPR compliance (column *P*). Of these, 6 approaches utilise semantic web to represent information about processes and activities, while other 7 approaches utilise other forms of information representations.
- **Ex-ante and Ex-post representations**: 12 approaches represent information in ex-ante phase (column *EA*), and 10 approaches represent information in ex-post phase (column *EP*). Of these, 6 approaches represent information in both phases. The other 13 approaches do not specify any indication of which phase they represent or do not consider phases in their representations.
- **Consent**: Only 9 approaches represent information about consent based on GDPR (column *C*), of which 3 do so using semantic web. This information concerns attributes associated with request for consent, context of given consent, and information about its withdrawal. Additionally, two approaches specifically mentioned in context of consent (Grando et al. and Consent Receipt) do not address GDPR but provide information representation relevant to consent.
- **Compliance Evaluation**: Evaluation of compliance is carried out by 18 approaches (column *Cm*), with 8 of those utilising semantic web in some form.
- **Suggestions and requirements**: 9 approaches provide requirements or suggestions to achieve compliance (column *Rq*). Of these, 7 approaches produce suggestions after an evaluation of compliance.
- **Open access to resources**: Of 29 approaches, only 4 approaches have published their resources in an open and accessible format (column *O*). The other resources either provide information on their work through publications or project deliverables, and do not provide a link to access referenced resources.

Following these observations, a more detailed description of analysis is provided for areas identified in Section 3.1. These correspond to investigation of state of the art regarding relevance to research objectives.

3.7.2 Representation of GDPR

The overview presented in Table 3.1 shows only three approaches representing structure of GDPR and enabling machine-readable linking of information to specific parts of its text. Of these, two (MIREL and DAPRECO) utilise PrOnto [41], [55] and are therefore consolidated into a single work for purposes of analysis. Therefore, there are only two distinct approaches for representation of GDPR in state of the art.

Table 3.2 presents an overview of these approaches, along with ELI ontology [42] used in publication of metadata associated with legislations including GDPR. The granularity of ELI ontology in representing metadata is limited to specifying information at document level, i.e. it defines GDPR as a legislation but does not provide metadata about its structure or contents. The EU Publication Office, which is in-charge of publishing legislations and maintains ELI ontology, has indicated[48] they are working on updates to ontology - these are represented in the table as 'ELI+' for descriptive purposes and completeness of analysis.

Table 3.2: Approaches representing GDPR in machine-readable format

Work	**ELI**	**ELI+**	**Agarwal et al**	**PrOnto**
Vocabulary	OWL2	OWL2	RDFS	Akoma Ntoso
Granularity	Legislation	Sub-Paragraph	Paragraph	Sub-Paragraph
Glossary	✗	✓	✗	✗
PID	✓	✓	✗	✗
OA	✓	✓	✗	✗
GDPR text	✗	✓	✗	✓

The table demonstrates that there are very few approaches working with machine-readable representations of GDPR, even when a large number of approaches utilise machine-readable formats for information associated with specifying information for compliance assessment. The table also shows that none of existing approaches address drawbacks of ELI, but instead utilise other vocabularies to achieve granularity regarding GDPR. For this, Agarwal et al. [40] use RDFS, while PrOnto uses Akoma Ntoso to represent hierarchy of text within GDPR. Apart from ELI, neither of other two resources are open and publicly accessible (row OA), and they do not use a persistent identifier (row PID) enabling sustained use over time. Also, no approach provides a glossary of terms utilised or defined by GDPR.

3.7.3 Representation of activities

The modelling of activities in SotA is elaborated in Table 3.4 in terms of its use in terms of phase (ex-ante or ex-post), concepts modelled, and representation. The column headings used are as follows - (Repr): representation of activities, (EA): Ex-ante modelling, (EP): Ex-post modelling, (Pu): Purpose, (Pr): Processing, (DS): Data Sharing, (Rp): Recipients, (St): Data Storage, (Rg): Rights, (LB): Legal Basis. The table uses a check mark (✓) to indicate that the approach represents a feature, while a blank denotes no information was found regarding its representation. The use of a blank is indicative of the possibility that a feature exists but is not published or accessible, or that it may be developed in future given ongoing work in these approaches.

[48] Information communicated via private channels and correspondences, and does not have a public reference.

The table shows use of BPMN notation and semantic web ontologies to model activities for GDPR, which includes reuse of existing ontologies PROV-O [47] and PWO [179]. The approaches model activities across both ex-ante and ex-post phases with 4 approaches modelling both phases.

Approaches modelling activities do not necessarily model their concepts to match or align with terminology utilised by GDPR. In cases where terminology is relevant to GDPR compliance, source of concepts and relationships used are not provided. Therefore, it is left up to an adopter's 'common sense' to interpret terms correctly, and to manually align such ontologies with other legal resources.

In approaches that utilise PROV-O and PWO to represent activities, developed ontologies extend existing ontologies by defining concepts relevant to GDPR. As PROV-O represents provenance information, and therefore records information about activities that have already been carried out (in the past), use of PROV-O to represent ex-ante activities involves recording information of planned activities as a provenance record. PROV-O defines the concept *Plan* to represent a plan of activities, but it is not utilised by any approach. P-Plan, an extension of PROV-O which further expands on *Plan* to define templates of activities as workflows, is also not utilised by any approach. PWO, by comparison, represents workflows and is therefore capable of expressing ex-ante information. However, the use of PROV-O provides commonality and interoperability by virtue of being a standardised vocabulary.

Table 3.4: Representation of activities in SotA

Work	Repr	EA	EP	Pu	Pr	DS	Rp	St	Rg	LB
SPECIAL	PROV-O	✓	✓	✓	✓	✓	✓	✓		
SPL+CitySPIN	PROV-O	✓	✓	✓	✓	✓	✓	✓		
MIREL	PWO	✓		✓	✓	✓	✓	✓	✓	
MRL+DAPRECO	PWO	✓		✓	✓	✓	✓	✓	✓	
BPR4GDPR		✓	✓	✓	✓	✓	✓			
Ujcich et al.	PROV-O		✓	✓	✓	✓	✓	✓	✓	✓
Lodge et al		✓		✓						
Tom et al	BPMN	✓		✓	✓	✓	✓	✓	✓	
LUCE		✓	✓			✓	✓			
Sion et al		✓		✓	✓	✓	✓	✓		✓
privacyTracker		✓	✓			✓	✓			
Basin et al		✓		✓						
RestAssured				✓	✓	✓	✓	✓		

Most approaches model some combination of purpose, processing, data item or category, sharing of data, recipients, and data storage within their representations. In comparison, representation of rights and legal basis associated with processing of personal data is not common with only a few approaches providing concepts for their representation. This is due to use of consent as an implicit legal basis in approaches with assumption that activities associated with rights are separate and distinct operations from processing activities.

3.7.4 Representation of Consent

Representation of consent in SotA is elaborated in Table 3.6, with the column headings as follows - (PD): Personal Data, (Pu): Purpose, (Pr): Processing, (Sh): Data Sharing, (St): Data

Storage, (Rp): Recipients, (S): Data Source, (W): Withdrawal of consent, (D): Delegation, (V): Visualisation, (SE): Significant effects of processing, (Ct): Context, (T): Type. The table uses a check mark (✓) to indicate that whether an approach represents that feature, while a blank denotes that no information was found regarding its representation. The use of a blank is indicative of possibility where that exists but is not published or accessible, or that it may be developed in future given ongoing work in these approaches. Some upcoming research projects have been included based on published information about their use and representation of consent.

Table 3.6: Representation of consent in SotA

Work	PD	Pu	Pr	Sh	St	Rp	S	W	D	SE	Ct	T
SPECIAL	✓	✓	✓	✓	✓	✓		✓				
SPL+CitySPIN	✓	✓	✓	✓	✓	✓		✓				
Lodge et al	✓	✓										
Peras	✓	✓	✓	✓	✓			✓				
Coletti et al	✓	✓					✓	✓				
AdvoCATE	✓	✓			✓	✓				✓	✓	
RestAssured	✓	✓	✓	✓	✓	✓						
OPERANDO	✓	✓	✓	✓		✓						
PoSEID-on	✓					✓						
MHMD	✓											
DECODE	✓	✓			✓							
Consent Receipt	✓	✓									✓	✓

In analysed approaches, there is no specific vocabulary to represent all attributes of consent as required for compliance with GDPR. Instead, aspects of given consent are represented in view of compliance requirements - such as within provenance of process flows or as a record of given consent. In this, all approaches define personal data involved, and most define purpose of processing.

However, overall, approaches lack metadata regarding consent as required to evaluate its validity based on GDPR. Few approaches explicitly define processing activities involved, data storage, data sharing and recipients. Only one approach enables representation of data source, and few approaches define right to withdraw consent. Furthermore, only one approach defines significant effects of processing as required to be provided for obtaining valid consent. Only two approaches capture context such as medium associated with consent. Only one approach provides types of consent (explicit or implicit). No approach considers possibility of delegation - such as when a parent or guardian provides consent in lieu of a minor/child.

Given consent, which includes information about choices made by data subject, is required to be recorded in order to demonstrate compliance with GDPR. Furthermore, the mechanism used to provide and obtain consent is also required to be recorded in order to assess its validity with requirements of given consent under GDPR. Therefore, process flows associated with provision of consent choices and obtaining consent are required to be recorded and can utilise the same mechanism as for defining process flows regarding processing activities. In this, existing approaches combine the two process flows by implicitly assuming legal basis of their processing activity to be consent, but do not capture context of given consent. In comparison, Consent Receipt is specifically designed to capture a record

of given consent as a receipt of transaction involving personal data. However, it does not contain required fields to express information about consent as required by GDPR.

From this, it is clear that existing state of the art does not provide sufficient means to represent information required to assess compliance of consent as defined by GDPR.

3.7.5 Querying of information associated with GDPR Compliance

The approaches described in state of the art do not provide guidelines or directions on querying of information as compared to representation of information and evaluation of compliance. Approaches that use RDF for information representation demonstrate or mention use of SPARQL for retrieval of information but do not provide evidence of practical applications of such queries being used to answer questions associated with compliance.

In comparison, approaches related to privacy policies focus on querying as part of their objective in retrieving relevant information from text of a privacy policy. In this, the UsablePrivacy project is notable in its use of SPARQL to retrieve information. Linden et al. demonstrate an approach that uses queries derived from GDPR-compliance checklists provided by ICO to identify coverage of GDPR within privacy policies [165].

Some research projects utilise legal experts to provide questions which are then translated into competency questions and used in development of ontologies and technology. Projects such as SPECIAL and MIREL demonstrate use of SPARQL queries derived from competency questions to retrieve information using developed ontologies. The details of this are sparse in their deliverables with some examples providing an insight into their use, but queries themselves are available for reuse. Additionally, none of existing approaches demonstrate an application of such queries to assist with compliance such as by adopting some authoritative document outlining compliance investigation process through queries or questionnaires.

From this, it can be summarised that while projects focus on automation of information representation and compliance evaluation, utilisation of querying to assist in compliance process has not seen significant development in legal compliance domain or has not been published in a reusable manner. Approaches involving privacy policies by comparison utilise querying to simplify information retrieval, its understanding, and demonstrate potential application towards legal compliance documentation.

3.7.6 Evaluation of GDPR Compliance

The evaluation of GDPR compliance in analysed approaches is summarised in Table 3.8, with column headings are as follows - (M): Method used for evaluation, (Sc): Scope of evaluation, (EA): Ex-Ante, (EP): Ex-post, (MR): Machine-readable outcome, (R): Suggests remedies and recommendations, (LG): Links results to GDPR. The table uses a check mark (✓) to indicate whether an approach represents that feature, while a blank denotes that no information was found regarding its representation. The use of a blank is indicative of a possibility that feature exists but is not published or accessible, or that it may be developed in future given ongoing work in these approaches. Some upcoming research projects have been included based on published information about their intended use of compliance evaluation methodologies.

Table 3.8: Compliance evaluation in SotA

Work	Method	Scope	EA	EP	MR	R	LG
SPECIAL	OWL	Consent	✓	✓	✓		
SPL+SERAMIS	ODRL	Obligations	✓		✓	✓	✓
SPL+Vos et al.	OWL, ASP	Obligations	✓		✓	✓	
SPL+CitySPIN	OWL	Consent	✓	✓	✓		
MIREL	RuleML	Obligations	✓	✓	✓	✓	✓
MRL+DAPRECO	RuleML	Obligations	✓	✓	✓	✓	✓
BPR4GDPR	OWL	Process Flows	✓	✓		✓	
Lodge et al	SDK	Process Flows	✓		✓	✓	
Tom et al	BPMN	Process Flows	✓		✓	✓	
Corrales et al	Questionnaire	Obligations	✓				
LUCE	Smart Contracts	Data Sharing	✓	✓	✓		
AdvoCATE	Smart Contracts	Consent		✓	✓		
Sion et al	UML, DFD	Process Flows	✓		✓	✓	
privacyTracker	Access Control	Data Sharing		✓	✓		
Robol et al	STS	Process Flows	✓		✓		
GuideMe	Questionnaire	Process Flows	✓			✓	
Basin et al	Algorithm	Process Flows	✓				
RestAssured	XACML	Process Flows	✓	✓	✓		
DEFeND	Questionnaire	Obligations	✓		✓		
OPERANDO	Access Control	Process Flows		✓	✓		
PoSEID-on	Smart Contracts	Data Sharing		✓	✓		
DECODE	Smart Contracts	Consent		✓	✓		

From the table, most approaches (9 of 22) focus on evaluation of process flows followed by obligations fulfilment (6 of 22) and given consent (3 of 22). In terms of method or formalism used, semantic web (OWL, RuleML) show significant utilisation (8 of 22). It can be summarised that approaches evaluate compliance in either ex-ante (17 of 22) and ex-post (12 of 22) phase, with some approaches covering both phases (7 of 22). Most approaches (18 of 22) produce a machine-readable outcome or artefact from the evaluation process that can be persisted for documentation purposes. Some approaches suggest remedial measures (9 of 22) to correct or rectify lapses in compliance based on evaluation. Only three (of 22) approaches link the evaluation and its outcomes to GDPR. Based on this, it can be concluded that the state of the art provides a variety of approaches and methodologies for evaluating GDPR compliance in ex-ante and ex-post phases, and features remedial measures to suggest corrections to achieve compliance. It is also evident that most outcomes of compliance evaluation are not associated or linked with aspects of GDPR - which is a gap in terms of how compliance information should be documented.

One noticeable absence in these approaches is the use of a validation mechanism such as SHACL or ShEx for RDF data validation. In approaches that utilise semantic web representations of information, use of SHACL - a W3C specification - would provide the means to express and ensure correctness of information, validate it using constraints, and persist results in RDF. SHACL can also be used only for verification of information before its evaluation for GDPR compliance, or conversely for checking of results post-evaluation to record a state of compliance. The absence of SHACL, and information validation in general, is therefore a noticeable gap within state of the art.

While most approaches record a machine-readable representation of compliance evaluation, there is a lack of work regarding how such information can be used for documentation

and demonstration of compliance. In approaches that link evaluations with GDPR, none explore the possibility of compliance coverage of GDPR carried out using recorded information. The links used in these approaches are either textual or based on interpretation of GDPR using developed ontologies - such as in PrOnto.

The source of compliance criteria within research projects are requirements and specifications provided by legal experts. The methodologies specified within these approaches do not provide access to specific queries or criteria fulfilled by its compliance evaluation. In case of research projects, evaluation of compliance is often driven by use-cases supplied by commercial partners which may restrict access to information deemed commercially sensitive. In approaches that produce recommendations or remedies from evaluation, it is unclear as to how these are recorded against evaluation and tracked in terms of changes made to processes across time - i.e. how compliance of a system evolves through these changes.

From this, it can be concluded that state of the art demonstrates use of technologies, and more specifically semantic web, in evaluation of GDPR compliance. It also demonstrates machine-readable artefacts as outcome of evaluations which are then used to suggest remedial measures to achieve compliance. These provide opportunities for utilising evaluation outcomes in documentation of ongoing compliance and documenting coverage of compliance GDPR. Furthermore, evaluation in these approaches is carried out under assumption of validity of existing data and does not consider absence of required information. This also provides an opportunity for assisting the compliance process by carrying out evaluations that check for validity of information in terms of existence and correctness.

3.8 Gaps and Opportunities for Further Work

The following list provides avenues for carrying out research based on opportunities identified from gaps within state of the art towards achieving the research objectives of this thesis:

1. **Machine-readable representation of GDPR:** A linked-data representation of GDPR based on existing ELI ontology with sufficient granularity to represent recitals and sub-paragraphs within articles. This will provide an ELI-compatible resource to link information with GDPR.
2. **Glossary of terms and concepts associated with GDPR:** A glossary of terms and concepts relevant to GDPR compliance that can provide interoperability between machine readable approaches, and using a machine-readable representation of GDPR to indicate source of information.
3. **Representation of activities associated with processing of personal data in ex-ante and ex-post phases:** A cohesive representation of activities across both ex-ante and ex-post phases that can indicate how an ex-ante activity acts as a plan for executing an ex-post activity. This will provide an indication of planned compliance in ex-ante stage, verification of compliant processing in ex-post stage, and evolution of compliance in a system across time based on changes to information in both phases. The PROV-O ontology provides a basis to represent activities using an standardised ontology based on similar use within state of the art.
4. **Representation of consent information:** An ontology for representing information associated with consent covering all requirements of GDPR. This can incorporate or

reuse representation of activities from above to indicate activities associated with consent.

5. **Demonstration using authoritative compliance queries:** Implementing retrieval of information for compliance using questions and checklists published by data protection offices to demonstrate practical application of research in compliance process.
6. **Validating information for compliance evaluation:** An approach to validate information to ensure its existence and correctness in order for it to be used in evaluating compliance. The approach needs to incorporate both ex-ante and ex-post phases, and can incorporate relationship between the two phases to demonstrate ongoing compliance. The use of SHACL is ideal for this purpose given that it is a semantic web standard and has the ability to persist its results in RDF which provide opportunity to query and use them for creating compliance documentation linked to GDPR.

These gaps directly influenced the establishment of research objectives presented in Section 1.2.2 in terms of using semantic web technologies for information representtion, querying, and validation for GDPR compliance.

This page intentionally left blank.

4 | Analysing GDPR Compliance Requirements

The analysis of state of the art in Chapter 3 provided identification of opportunities for addressing the gaps within it. These opportunities relate to research objective $RO3$ regarding construction of ontologies, $RO4$ for querying, and $RO5$ for information validation. In order to achieve these, it is imperative to have an understanding of GDPR and its compliance requirements as required by research objective $RO1$. The requirement gathering process is shaped by the scope of research question - which for this thesis is representation of activities associated with processing of personal data and consent. The identified requirements then need to be expressed in a form which will facilitate representation of information as ontologies, its querying, and validation in compliance process. This is required to fulfil research objective $RO2$.

The approaches presented in SotA in Chapter 3 directly delve into obligations and requirements of data controllers to demonstrate compliance or towards data subject's consent and rights. Since requirements of GDPR compliance are also influenced by other stakeholders such as processors and authorities which play an unspecified role in the context of information associated with compliance, such roles consist of interactions between stakeholders and involve communication of information such as instructions provided by a data controller to a processor. Therefore, along with requirements of compliance, it is also important to understand interactions between stakeholders, information involved in such interactions, and requirements of information in terms of its interoperability between them.

As an example, consider a data controller that can have any number of internal representations of information necessary to demonstrate compliance, but an investigation by a supervisory body requires such information to be provided as per their stated requirements in an mutually understandable form. Furthermore, a processor contracted by a controller may also be required to present information relevant in investigation of compliance - which would also need to be mutually understandable by the processor, controller, and investigating authority. When utilising technological solutions for management of compliance information, analysis of information interoperability requirements enable identifying applications of such solutions within a larger context comprised of multiple stakeholders involved in the compliance process.

This chapter therefore first presents an analysis of GDPR in terms of stakeholders and interoperability of information between them to construct a model of information interoperability in Section 4.1. The model provides an analysis of requirements in terms of information and interoperability from the perspective of interactions between entities. This

provides an overview of information requirements which is used to find additional applications for existing information representation approaches and to identify gaps which can be addressed through future opportunities. Such analyses and requirements gathering related to GDPR also benefit the larger community and domain by providing information for standardisation activities in understanding role of information and its interoperability between stakeholders, such as those for DPVCG (see Section 1.4.6).

Following the above, Section 4.2 frames 'compliance questions' that provide information requirements necessary to evaluate compliance, with its methodology presented in Section 4.2.1, and Section 4.2.3 presenting assumptions and constraints that can be used to validate information for correctness and completeness. The use of compliance questions as competency questions in development and evaluation of ontologies is presented in Chapter 5, and use of constraints for validations is presented in Chapter 6.

4.1 Interoperability Model of Information based on GDPR

This section presents a model of interoperability for information associated with stakeholders based on an analysis of GDPR in terms of interactions between stakeholders and information involved in such interactions. The model enables understanding role of stakeholders in compliance process in terms of information requirements and provides a framework to establish relationship between interactions and information required for compliance. The model also provides motivation to incorporate interoperability as a core requirement within representations of information towards GDPR compliance. This provides context for development of ontologies presented in this thesis in terms of their intended application and usefulness to stakeholders associated with GDPR compliance.

The model serves to place contributions presented in this thesis within the larger context of stakeholders involved in GDPR compliance process. It guides the design and impact of ontologies presented in Chapter 5 in terms of establishing interoperability as a requirement based on potential exchange of information between stakeholders. Additionally, it also provides context about the roles and activities of stakeholders regarding information management and documentation for GDPR compliance.

The work described in this section was published within the interoperability and standardisation community as a conference paper [75] which was later expanded upon in a journal article [67] and a book chapter [68].

Interoperability Model

The creation of a model is based on identifying categories of entities as defined within GDPR and identifying interactions between them. Figure 4.1 visualises interactions between entities, and consists of Data Subject (DS), Data Controller (DC), Data Processor (DP) and Supervisory Authority[1] (SA) as entities defined within GDPR. The points of interactions consist of potential information exchange between entities and are guided by requirements of compliance. For example, interaction between data subject and data controller consists of data subject providing personal data to controller, while the controller is required to provide a copy of provided personal data for fulfilment of rights granted by GDPR.

[1] Supervisory Authority are also referred to as Data Protection Commission or Regulatory Body

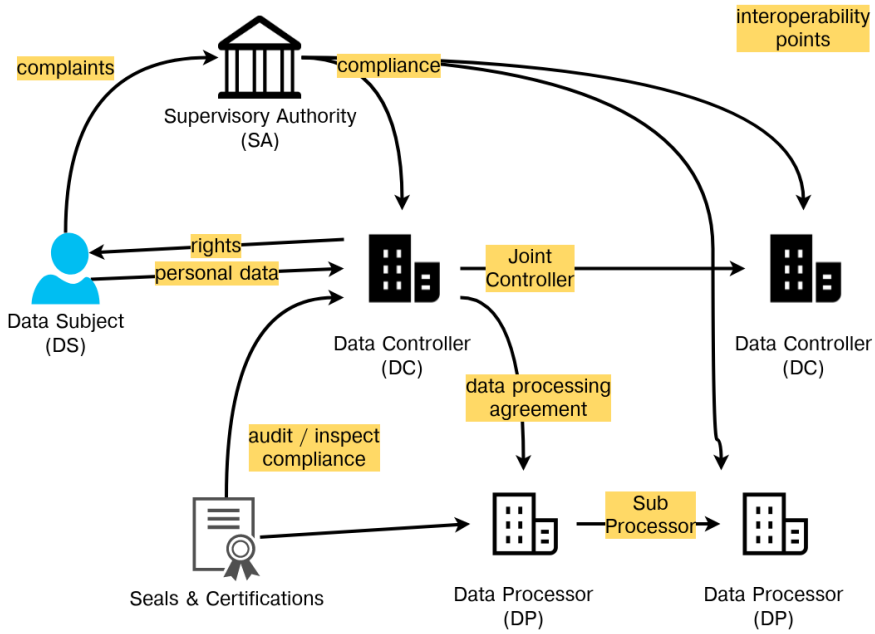


Figure 4.1: Interactions between entities based on requirements of GDPR [67]

The model and information categories serve to provide a larger context of information requirements of GDPR and demonstrate areas for application of contributions presented in this thesis. They also provide future development and application of presented work to other areas of GDPR compliance. More specifically, they provide opportunities where developed work can be extended or re-applied to represent additional information regarding activities associated with GDPR compliance - such as for data breach records, carrying out rights, recording compliance procedures, or data processing agreements.

Information Categories

The analyses of information flows between entities (additional description presented in publications [67], [75]) provides categorises of information requirements as: provenance records, consent information, data processing agreements, compliance information, and seals/certifications - of which provenance records and consent have a direct bearing on the stated research objectives. Carrying out further investigation and analyses of interactions and information involved is out of scope for research presented in this thesis given the focus of research question on information about activities associated with processing of personal data and consent. To this end, analyses of information categories presented below concerns only categories of provenance and consent, and provides requirements for design of ontologies required by research objective $RO3$.

Provenance records

Provenance in this case refers to information about entities and activities involved in the compliance process where a record is required to be kept and exchanged for compliance purposes. For example, GDPR requires controllers and processors to maintain provenance records of processing activities carried out under their responsibility in order to maintain and demonstrate compliance to supervisory authorities. Provenance records are also re-

quired to be maintained to enable provision of rights to data subjects, and for information sharing between controllers and third parties.

The information stored within provenance records is related to demonstrating compliant processing of personal data and fulfilment of obligations towards demonstration of compliance. They are modelled within state of the art (see Chapter 3) variously as logs, life-cycles, workflows, activities, and process flows. The term provenance in this case refers to both ex-ante and ex-post phases and is indicative of provenance of information to specify its existence. Therefore provenance in ex-ante phase means record of a model or plan used to indicate future processing of personal data, while that in ex-post phase refers to record or log of processing.

Since provenance information as described above encompasses information about artefacts and processes related to compliance, sharing this information in an interoperable format with other entities benefits both in their obligations regarding compliance. For example, controllers and supervisory authorities sharing interoperable compliance documentation can use the same set of tools and approaches for interacting with information. Also, by recording processes associated with compliance as provenance records, the same interoperable methods can be used to maintain, document, and demonstrate compliance. This is especially useful where information is to be shared between joint controllers and processors that need to exchange plans of processing for agreement as well as logs for successful implementations.

Since a data controller is required to ensure their intended activities are compliant as well as maintain records of processing activities regarding use of personal data, provenance records to be documented consist of intended plan of processing activities and their execution. Similarly, provenance records of consent requests and provision also need to be maintained to demonstrate compliance. Given the similarity in requirements for maintaining provenance records for different reasons, it is beneficial to provide a singular interoperable representation of provenance activities in a form that can capture and represent all aspects of provenance records required for compliance.

Consent information

As per GDPR, consent is an assent or agreement by a data subject regarding processing of their personal data for specified purposes by one or more entities. For compliance, a controller must record information regarding how consent was requested and choices provided by a data subject as given consent. GDPR has several obligations and requirements when it comes to valid consent, with additional requirements depending on sensitivity of personal data and processing. Artefacts associated with choices offered for consent therefore need to be preserved to demonstrate the validity of consent obtained using those choices. Similarly, consent revocation or revisions also need to be recorded and linked to earlier instances of consent to demonstrate a change was valid as per requirements of GDPR.

The processes associated with offering consent choices, retrieving and recording given consent, enabling withdrawal of consent, and executing revocation in internal processes need to be documented for compliance by relevant controller(s). These are associated across both ex-ante and ex-post phases where consent choices, provision of right to withdrawal, and demonstrating utilisation of consent in processing are demonstrated as ex-ante mea-

sures while given consent and revoked consent are ex-post artefacts.

If consent is considered as an instance of personal data, the same information model used to document processing of personal data can be re-purposed or reused to represent activities associated with consent. In addition to this, activities need to capture different stages of consent and personal data by representing their life-cycles which involve processes and artefacts which use them or are dependant on them to indicate their evolution and use over time.

Stating interoperability as a requirement

The model and analyses of information flows between entities provides motivation for establishing interoperability in representation of information to be exchanged. As the scope of thesis focuses on representation of activities associated with processing of personal data and consent, requirements gathered from this analysis relate primarily to information associated with provenance of activities and information about consent. Within this scope, ensuring potential reuse towards other categories through future extensions is provided by developing an interoperable ontology by using standards that can be utilised to also represent data processing agreements and compliance agreements in the future. This involves using standards of RDF and OWL2 to represent ontologies along with reuse of existing standardised vocabularies such as PROV-O and ELI. In addition, the research also provides transparency by using terminology of GDPR in developed ontologies, indicating requirements used to shape design of ontology and indicating source of concepts within GDPR.

Representing provenance is not limited to representation of processing of personal data, but is also applicable to information about other categories - consent, data processing agreements, compliance, and certifications. Therefore, utilising the same or compatible representation in representing provenance across all use-cases has advantage towards cohesive management of all associated information for compliance - and provides an objective for future work in expanding contributions presented in this thesis.

With this motivation, the next parts of this chapter provide requirements gathered for representing information regarding processing of personal data and consent while also including other relevant information such as data breaches and provision of rights to indicate potential applicability and reuse of developed ontologies in representing information through provenance records for GDPR compliance.

4.2 Compliance Questions

This section presents 'compliance questions' whose answers provide information necessary for evaluating GDPR compliance. The questions are essential to development of information representations within compliance management systems by providing requirements for structuring of information and its validation. Within this thesis, they are used to guide development of ontologies as competency questions (see Chapter 5) and validation of information (see Chapter 6). **The questions presented here are by no means exhaustive but represent gathered requirements from authoritative sources.** As supervisory authorities and courts continue to clarify and interpret compliance requirements of GDPR, these questions are expected to change and expand in future.

4.2.1 Methodology

The questions were created by studying authoritative sources regarding GDPR compliance consisting of data protection commissions, legal experts and agencies - that have published guidelines and resources to assist organisations with the process of establishing and maintaining GDPR compliance. In this process, each identified clause or article of GDPR pertaining to the research question was formulated as a question, with the above mentioned sources providing indication on how the question should be interpreted and requirements for its compliance. The questions were derived from reading and understanding of compliance requirements and are intentionally expressed as a 'simple question' whose answering requires minimal information in order to determine requirements of such information towards constructing an ontological representation of it.

For questions presented in this thesis, following sources were used or referenced in addition to text of GDPR:

- Guidelines, clarifications, and discussions on interpretation of GDPR published by European Data Protection Board[2] (EDPB)
- Guidelines, clarifications, and discussions on interpretation of GDPR published by Article 29 Working Party[3] and endorsed by EDPB
- Resources published Data Protection Commission[4] (Ireland) - with particular focus on document 'GDPR guidance for SMEs'[5]
- Resources published by Information Commissioners Office[6] (United Kingdom), with particular use of 'Data protection self assessment for organisations'[7]
- Resources published by federated data protection offices in Germany, in particular the audit checklist published by Lower Saxony Data Protection Authority[8] self-translated from German to English[9]
- Resources regarding GDPR compliance published by professional institutions within legal compliance domain, specifically - Nymity[10] and IAPP[11]
- Executive and Court decisions regarding GDPR compliance, tracked through an online community service https://www.enforcementtracker.com/ which also denotes relevant articles of GDPR

The questions pertaining to consent and its associated processing activities were validated through consultation with a law professor at Trinity College Dublin who served as a legal domain expert and validated the questions and information required to answer them. This consisted of providing a spreadsheet containing categories of questions along with their assumptions and constraints as information requirements, with the feedback consisting of whether stated information was correct, utilised correct terminology, and changes in statements to suit legal interpretations. Other questions pertaining to processor obligations, data

[2] https://edpb.europa.eu/
[3] https://ec.europa.eu/newsroom/article29/news.cfm?item_type=1360
[4] https://dataprotection.ie/
[5] https://dataprotection.ie/en/guidance-landing/guidance-smes
[6] https://dataprotection.ie/
[7] https://ico.org.uk/for-organisations/data-protection-self-assessment/
[8] https://lfd.niedersachsen.de/download/146715/
[9] https://doi.org/10.5281/zenodo.3380469
[10] https://info.nymity.com/gdpr-compliance-toolkit
[11] https://iapp.org/resources/article/gdpr-genius/

breach requirements, and rights were not validated through this process due to unavailability of expert and their lower priority in relevance to the research question.

The author of this thesis was part of an informal consultation on a real-world (details confidential) research project at MasterCard regarding creation of a semantic web ontology for expressing consent information based on requirements of the GDPR. The consultation consisted of identifying information for representing consent based on requirements of GDPR and designing an ontology to represent it. This process provided real-world requirements about management of consent information which had an influence on compliance questions associated with consent and design of GConsent ontology presented in Section 5.4.

The collected questions were rephrased and divided into smaller more granular questions towards establishment of information requirements and constraints. Each question was assigned an ID to enable tracking its use. Where possible, each question was associated with specific clauses of GDPR by denoting an article or recital relevant to it. Each question was analysed in terms of information requirements to identify assumptions that clarify interpretation of the question, and constraints that express a condition that can be tested and used to validate information. For each identified constraint, a failing test case was identified that did not satisfy the condition and could be used to test its validation.

The list of questions is presented in Section 4.2.2 with assumptions and constraints presented in Section 4.2.3. Along with compliance questions regarding activities associated with processing of personal data and consent, the list also contains questions about additional activities associated with right to be informed and reporting of data breach by controllers. These questions are used in development of ontologies that demonstrate application of a common approach to represent activities related to processing of personal data, consent, data breach, and provision of rights for GDPR compliance.

4.2.2 List of Questions

The questions are categorised based on the topic they focus on within the context of compliance. Where a question was directly derived from specific clause(s) a reference to the clause(s) is provided at end of sentence[12]. In order to provide a context for application of questions within real world, a group of use-cases are provided in Section 4.2.2.1. An overview of compliance questions and their categorisation as follows:

- Section 4.2.2.2 Compliance questions pertaining to records of processing activities to be kept by controllers: `CMQ.1 - CMQ.15`.
- Section 4.2.2.3 Compliance questions pertaining to records of processing activities to be kept by processors: `CMQ.16 - CMQ.25`.
- Section 4.2.2.4 Compliance questions pertaining to legal basis of processing activities: `CMQ.26 - CMQ.27`.
- Section 4.2.2.5 Compliance questions pertaining to personal data: `CMQ.28 - CMQ.34`.
- Section 4.2.2.6 Compliance questions pertaining to given consent: `CMQ.35 - CMQ.69`.
- Section 4.2.2.7 Compliance questions pertaining to change in consent state: `CMQ.70 - CMQ.87`.

[12]Shortened references are used to indicate the clause of GDPR. For example: R32 indicates Recital 32, A7-2 indicates Article 7 Para 2, and A9-2c indicates Article 9 Para 2 Sub-Para c.

- Section 4.2.2.8 Compliance questions pertaining to provision of right to be informed: `CMQ.88 - CMQ.105`.
- Section 4.2.2.9 Compliance questions pertaining to reporting of data breach by controllers: `CMQ.106 - CMQ.120`.

4.2.2.1 Use-cases

To identify application of compliance questions in various scenarios within real world, 15 categories of use-cases were identified to determine necessary information required to identify their ontological representations. The use-cases are based on identifying information affecting compliance requirements and identifying variances which affect representations of that information. These were useful to guide development of ontologies by serving as test scenarios where an ontology could be applied and evaluated for representing information. These use-cases are not intended to be comprehensive or normative from a legal compliance point of view, but are helpful in understanding presented work by providing a context for their application.

A summary of these use-cases is as follow:

1. Obtaining / Declaring Consent (its state)
   (a) The consent is given
   (b) Consent was given, but is now invalidated (by the controller)
   (c) Consent was given, but was withdrawn (by the Data Subject)
   (d) Consent was requested (by the controller)
   (e) Consent was requested, but was refused (by the Data Subject)
   (f) Consent state is unknown (e.g. when importing data about consent)
2. Entity the consent is about
   (a) The consent is about a Data Subject who is not a minor
   (b) The consent is about a Data Subject who is a minor
3. Activity for Data Subject
   (a) There was an age verification process associated with the consent (such as for minors)
   (b) There was an identity verification process associated with the consent
4. Entity that provided consent
   (a) Consent was provided by the Data Subject it is about
   (b) Consent was not provided by the Data Subject it is about, but was provided by a Delegation
      i. Consent in the Delegation was provided by another Data Subject
      ii. Consent in the Delegation was provided by a Person
      iii. Consent in the Delegation was provided by another Delegation
5. Role within Delegation
   (a) Entity is the Parent/Guardian of the Data Subject
   (b) Entity is a third-party to the Data Subject
6. Activity of Delegation
   (a) There was some verification process to assert the authentication of the delegation
7. Personal Data associated with consent
   (a) Consent was given for specific instances of personal data e.g. John/Jane Doe

(b) Consent was given for categories of personal data e.g. Name

8. Medium of Consent
   (a) consent is given via a web-form
   (b) consent is given as a signed paper document
   (c) consent is given as a verbal confirmation
   (d) consent is given implicitly in some form (medium)
   (e) consent is given via delegation in some form (medium)
9. Activity responsible for consent
   (a) Activity created consent as a new entity
   (b) Activity modified existing consent
10. Previous consent and relationship
   (a) Consent has no previous instance
   (b) Consent has a previous instance and replaces it
11. Differences between consent instances
   (a) Something changes between two consent instances (e.g. personal data category is added)
12. Time constraints
   (a) consent expires (has a tangible expiry such as a specific date or duration)
   (b) consent does not expire (is valid for "as long as required")
13. Third party Association
   (a) Personal Data is collected from a third party
   (b) Personal Data is stored with a third party (processor)
   (c) Personal Data is shared with a third party
   (d) Processing involves third party
   (e) Purpose involves third party
14. Role of Third Party
   (a) Third Party is a Processor contracted by the Controller
   (b) Third Party is another Controller
   (c) Third Party is another entity (regulatory/supervisory/governmental)
15. Storage Duration and Locations for Personal Data
   (a) Data is stored for a fixed time (specific instance or duration)
   (b) Data is stored for an indefinite duration ("for as long as required")

4.2.2.2 Compliance questions pertaining to records of processing activities to be kept by controllers

CMQ.1 How are the records of processing activities maintained? (R82,A30,A30-3)

CMQ.2 What is the name and identity of the controller(s) and their representatives/DPOs? (A30-1a)

CMQ.3 What are the purposes of processing? (A30-1b)

CMQ.4 What are the categories of data subjects? (A30-1c)

CMQ.5 What are the categories of personal data? (A30-1c)

CMQ.6 Is data shared?

CMQ.7 If data is shared, what are the categories of recipients to whom the personal data is or will be disclosed? (A30-1d)

CMQ.8 If data is shared, what are the identities of the recipients to whom the data is or will be disclosed? (A30-1d,A30-1e)

CMQ.9 If data is shared, are the recipients to whom the data is or will be disclosed based in a Third Country or International Organisation? (A30-1e)

CMQ.10 If data is shared, and the recipients are in a Third Country or International Organisation, what are the safeguards associated with data transfer? (A30-1e)

CMQ.11 Is data stored?

CMQ.12 Where data is stored, its erasure is based on what criteria: time limit or condition or event?

CMQ.13 Where data is stored, what are the time limits or conditions or events for erasure for different categories of data? (A30-1f)

CMQ.14 What are the technical and organisational security measures w.r.t to the processing of personal data? (A30-1g)

CMQ.15 Where data is shared, what are the purposes for sharing of personal data with the recipients?

4.2.2.3 Compliance questions pertaining to records of processing activities to be kept by processors

CMQ.16 How are the records of processing activities maintained? (R82,A30,A30-3)

CMQ.17 What is the name and identity of the processor(s) and their representatives/DPOs? (A30-1a)

CMQ.18 What is the name and identity of the controller(s) the processor is acting on behalf of? (A30-2a)

CMQ.19 What are the categories of processing carried out on behalf of the controller? (A30-2b)

CMQ.20 If data is shared, what are the categories of recipients to whom the personal data is or will be disclosed? (A30-1d)

CMQ.21 If data is shared, what are the identities of the recipients to whom the data is or will be disclosed? (A30-1d,A30-1e)

CMQ.22 If data is shared, are the recipients to whom the data is or will be disclosed based in a Third Country or International Organisation? (A30-1e)

CMQ.23 If data is shared, and the recipients are in a Third Country or International Organisation, what are the safeguards associated with data transfer? (A30-1e)

CMQ.24 What are the technical and organisational security measures w.r.t to the processing of personal data? (A30-1g)

CMQ.25 Where data is shared, what are the purposes for sharing of personal data with the recipients?

4.2.2.4 Compliance questions pertaining to legal basis of processing activities

CMQ.26 What is the legal basis for processing of data?

CMQ.27 What is the legal basis for the purpose for processing of data?

4.2.2.5 Compliance questions pertaining to personal data

CMQ.28 What are the sources of personal data?

CMQ.29 What personal data are collected from the data subject?
CMQ.30 Where personal data are not collected from the data subject, what are their sources?
CMQ.31 Where data has been anonymised, what techniques were used for anonymisation?
CMQ.32 Can pseudo-anonymised data be de-anonymised by the organisation using information it already possesses or is available to it?
CMQ.33 Where personal data is collected, is it pseudo-anonymous?
CMQ.34 What are the special categories of personal data being processed?

4.2.2.6 Compliance questions pertaining to given consent

CMQ.35 Who is the Data Subject associated with consent? (A4-11)
CMQ.36 What are the Personal Data associated with consent? (R32,A4-11)
CMQ.37 What are the Purposes associated with consent? (R32,R42)
CMQ.38 What are the Data Processing associated with consent? (R32,A4-11)
CMQ.39 What is the current Status of consent? (A7-3)
CMQ.40 Who are the Data Controllers associated with consent?
CMQ.41 Who provided consent? (A7-2)
CMQ.42 Was consent provided by Delegation? (A8-c)
CMQ.43 If consent was provided by Delegation, what was the role played by Delegate with respect to the Data Subject?
CMQ.44 If consent was provided by Delegation, how was the delegation executed?
CMQ.45 If consent was provided by Delegation, how was the delegate authenticated? (A8-2)
CMQ.46 Who was the consent given to?
CMQ.47 If consent was not given to the Data Controller, what is the relationship between the entity it was provided to and the Data Controller?
CMQ.48 How was the consent given/obtained?
CMQ.49 What artefacts were involved in the giving/obtaining of consent?
CMQ.50 What were the choices provided for consent?
CMQ.51 What was the statement or affirmative action indicating given consent?
CMQ.52 How was the right to withdraw consent communicated to the data subject?
CMQ.53 At what location was the consent given?
CMQ.54 What is the medium associated with consent? (R32,A7-2)
CMQ.55 What is the timestamp associated with the consent?
CMQ.56 What is the expiry of the consent?
CMQ.57 Is the purpose or processing associated with a third party?
CMQ.58 What is the role played by the third party in the purpose or processing?
CMQ.59 Does the processing of data involve storage of data?
CMQ.60 If personal data is being stored, what is the duration of storage for Personal Data?
CMQ.61 If personal data is being stored, what is the location of storage?
CMQ.62 Are processing associated with consent of automated nature? (R71,A9-2c,A22-2c)
CMQ.63 Does the processing of data involve transfer to a Third Country or International Organisation? (R111,A49-1a)
CMQ.64 If processing of data involves transfer to a Third Country or International Organisation, what is the identity of the Third Country or International Organisation?
CMQ.65 Do the personal data associated with consent belong to a special category? (R51,A8-2a)

CMQ.66 How is personal data associated or linked to the data subject?
CMQ.67 Is the Data Subject of legal age to provide their own consent? (A8)
CMQ.68 What are the specific laws that determine the legal age to provide consent? (A8-1)
CMQ.69 Does the Data Subject have a specific relationship with the Data Controller? (R43)

4.2.2.7 Compliance questions pertaining to change in consent state

In this, the definition of change in consent is where the state/status of consent is changed. Example: unknown to not asked or not given, from given to withdrawn, from given to invalidated. Changes where the result is given consent or obtained consent where none existed is considered as given consent with compliance questions listed in the previous section.

CMQ.70 Who is the Data Subject associated with consent? (A4-11)
CMQ.71 What are the Personal Data associated with consent? (R32,A4-11)
CMQ.72 What are the Purposes associated with consent? (R32,R42)
CMQ.73 What are the Data Processing associated with consent? (R32,A4-11)
CMQ.74 What is the current state/status of consent? (A7-3)
CMQ.75 Who are the Data Controllers associated with consent?
CMQ.76 Who changed the state/status of consent?
CMQ.77 Was consent changed by Delegation?
CMQ.78 If consent was changed by Delegation, what was the role played by Delegate with respect to the Data Subject?
CMQ.79 If consent was changed by Delegation, how was the delegation executed?
CMQ.80 If consent was changed by Delegation, how was the delegate authenticated? (A8-2)
CMQ.81 How was the consent state/status changed?
CMQ.82 What artefacts were involved in the change in state/status of consent?
CMQ.83 If change in consent was done by the Data Subject, what was the statement or affirmative action indicating change to their consent?
CMQ.84 At what location was the consent changed?
CMQ.85 What is the medium associated with change in consent? (R32,A7-2)
CMQ.86 What is the timestamp associated with the consent?
CMQ.87 If the current state/status of consent is valid for processing, what is the expiry of the consent?

4.2.2.8 Compliance questions pertaining to provision of right to be informed

CMQ.88 How was information relevant for the right to be informed provided to the data subjects?
CMQ.89 When was the information relevant to right to be informed was provided to the data subject?
CMQ.90 Was the name and contact details of the controller's representative provided to the data subject under the right to be informed?
CMQ.91 Was the name and contact details of the DPO provided to the data subject under the right to be informed?
CMQ.92 Was the purposes for processing provided to the data subject under the right to be informed?

CMQ.93 Was the legal basis for processing provided to the data subject under the right to be informed?

CMQ.94 Where the legal basis for processing was legitimate interest, was this communicated to the data subject under the right to be informed?

CMQ.95 If personal data is not obtained from the data subject, were the categories of personal data obtained communicated to the data subject under the right to be informed?

CMQ.96 If personal data is not obtained from the data subject, were the sources of data communicated to the data subject under the right to be informed?

CMQ.97 Where personal data is shared, were the recipients or categories of recipients communicated to the data subject under the right to be informed?

CMQ.98 If personal data is transferred to a third country or international organisation, were the identity of the third country or international organisation communicated to the data subject under the right to be informed?

CMQ.99 Where personal data is stored, were the retention period communicated to the data subject under the right to be informed?

CMQ.100 Were the rights available communicated to the data subject under the right to be informed?

CMQ.101 If the data subject provided consent, was the right to withdraw consent provided under the right to be informed?

CMQ.102 Was the right to lodge a complaint with a supervisory authority provided to the data subject under the right to be informed?

CMQ.103 Where personal data needs to be provided under statutory or contractual obligation, was this communicated to the data subject under the right to be informed?

CMQ.104 Where personal data needs to be provided under statutory or contractual obligation, and if this data needs to be obtained from the data subject, was this communicated to the data subject under the right to be informed?

CMQ.105 Where automated-decision making, including profiling is used, was this communicated to the data subject under the right to be informed?

4.2.2.9 Compliance questions pertaining to reporting of data breach by controllers

CMQ.106 When did the data breach occur?

CMQ.107 When did the controller become aware of the data breach? (R85,R33-1)

CMQ.108 Was the data breach notified to the supervisory authority?

CMQ.109 When was the notification of data breach provided to the supervisory authority? (R85)

CMQ.110 How was the notification of data breach provided to the supervisory authority? (R85,R33-1)

CMQ.111 Is the data breach likely to result in a high risk to the rights and freedoms of the natural person whose data is associated with it? (R86,A33-1,A34-1)

CMQ.112 Who are the data subjects whose personal data are associated with the data breach?

CMQ.113 Was the data breach notified to the data subjects?

CMQ.114 When was the notification of data breach provided to the data subjects?

CMQ.115 How was the notification of data breach provided to the data subjects? (R85,R33-1)

CMQ.116 How did the notification of data breach to the data subject provide information about the data breach? (R86,A34-2)

CMQ.117 How did the notification of data breach to the data subject provide information about mitigating potential effects? (R86,A34-2)

CMQ.118 What data was involved in the data breach?

CMQ.119 What technical measures were in place for the protection of data involved in the data breach? (R88)

CMQ.120 What steps were taken to prevent or mitigate the effects of the data breach?

4.2.3 Assumptions & Constraints

An assumption is defined as information or condition that always holds true and is useful in the interpretation of the compliance question. A constraint is defined as a condition that the information pertaining to compliance question must satisfy in order for it to be valid. The assumptions and constraints are listed with reference to the relevant compliance question at the end in brackets. An overview of the assumptions and constraints based on the categories of compliance questions is as follows:

- Section 4.2.2.2 Compliance questions pertaining to records of processing activities to be kept by controllers: `CMQ.1 – CMQ.15` has Assumptions `A.1 – A.5` and Constraints `C.1 – C.18`
- Section 4.2.2.3 Compliance questions pertaining to records of processing activities to be kept by processors: `CMQ.16 – CMQ.25` has Constraints `C.16 – C.32`
- Section 4.2.2.4 Compliance questions pertaining to legal basis of processing activities: `CMQ.26 – CMQ.27` has Assumptions `A.6` and Constraints `C.33 – C.34`
- Section 4.2.2.5 Compliance questions pertaining to personal data: `CMQ.28 – CMQ.34` has Assumptions `A.7` and Constraints `C.35 – C.38`
- Section 4.2.2.6 Compliance questions pertaining to given consent: `CMQ.35 – CMQ.69` has Assumptions `A.8 – A.28` and Constraints `C.39 – C.69`
- Section 4.2.2.7 Compliance questions pertaining to change in consent state: `CMQ.70 – CMQ.87` has Assumptions `A.30 – A.39` and Constraints `C.70 – C.87`
- Section 4.2.2.8 Compliance questions pertaining to provision of right to be informed: `CMQ.88 – CMQ.105` has Constraints `C.88 – C.113`
- Section 4.2.2.9 Compliance questions pertaining to reporting of data breach by controllers: `CMQ.106 – CMQ.120` has Constraints `C.114 – C.130`

4.2.3.1 Assumptions

A.1 Processing activities as a whole can involve multiple controllers - (CMQ2)

A.2 Data recipient categories refer to a collection or abstraction of recipients based on some context such as purpose e.g. “our ad partners” or requirement e.g. “law agencies” - (CMQ8)

A.3 Data recipients cannot be a collective or a group whose identity cannot be represented as a list of its members e.g. “our partners” vs “our partners – A, B, C” - (CMQ9)

A.4 Identity of the recipients does not need to be associated with the processing record if category of recipients is already documented. However, the identity of recipients within the category at that point of time must be recorded (somewhere). - (CMQ10)

A.5 Where data is stored, its erasure can depend on a condition or an event e.g. “XX days after your last sign-in”, or “as long as your account is active” - (CMQ12)

A.6 Processing carried out for a specific purpose adopts the legal basis of the purpose - (CMQ26)

A.7 Personal data may not have the source information associated with it directly. It must have some link (chain, path) to the information and its source from which it was obtained. - (CMQ28)

A.8 If there are multiple categories of personal data, consent is granted for all (union) of them - (CMQ36)

A.9 If a consent is given for multiple purposes, consent is considered given for all (union) of them - (CMQ37)

A.10 If a consent is given for multiple processing, consent is considered given for all (union) of them - (CMQ38)

A.11 Valid status of consent are when it is given (explicitly or implicitly) by the data subject, or by delegation - (CMQ39)

A.12 Invalid status of consent are when it its status is unknown, refused, not offered, withdrawn, invalidated, terminated, or expired. - (CMQ39)

A.13 The status of consent indicates whether it can be used as a legal basis for processing - (CMQ39)

A.14 Consent provided by a Person that is not the Data Subject is consent by Delegation - (CMQ41)

A.15 A delegation can involve another delegation for the provision of consent - (CMQ42)

A.16 If consent is provided to an actor not the data controller associated with consent, the actor is considered as acting on behalf of the controller - (CMQ46)

A.17 Specifying location for obtained consent is optional - (CMQ53)

A.18 Specifying medium for obtained consent is optional - (CMQ54)

A.19 Consent may not have a tangible expiry - (CMQ56)

A.20 Consent may have multiple forms of expiry depending on conditions or events - (CMQ56)

A.21 A purpose or processing may be associated with zero or more third parties - (CMQ57)

A.22 Processing of data may involve storage of data - (CMQ59)

A.23 Different personal data, processing, or purpose may have different storage of data - (CMQ60)

A.24 Storage duration may not be a tangible instance in time, it can depend on conditions or event - (CMQ60)

A.25 Processing may involve transfer of data to a third country or international organisation - (CMQ63)

A.26 Personal data associated with consent may belong to a special category - (CMQ65)

A.27 A data subject may be a minor or a child - (CMQ67)

A.28 The data subject may have a relationship of relevance with the Data Controller - (CMQ69)

A.29 If there are multiple categories of personal data, consent is granted for all (union) of them - (CMQ71)

A.30 If a consent is given for multiple purposes, consent is considered given for all (union) of them - (CMQ72)

A.31 If a consent is given for multiple processing, consent is considered given for all (union) of them - (CMQ73)

A.32 Valid status of consent are when it is given (explicitly or implicitly) by the data subject,

or by delegation - (CMQ74)

A.33 Invalid status of consent are when it its status is unknown, refused, not offered, withdrawn, invalidated, terminated, or expired. - (CMQ74)

A.34 The status of consent indicates whether it can be used as a legal basis for processing - (CMQ74)

A.35 Consent state/status can be changed by Delegation - (CMQ76)

A.36 A delegation can involve another delegation for the provision of consent - (CMQ77)

A.37 Specifying location for changed consent is optional - (CMQ84)

A.38 Specifying medium for changed consent is optional - (CMQ85)

A.39 Consent may not have a tangible expiry - (CMQ87)

A.40 Consent may have multiple forms of expiry depending on conditions or events - (CMQ87)

4.2.3.2 Constraints

C.1 Processing carried out by a controller must have information on how its records are being maintained - (CMQ1)

C.2 Records of processing activity under the responsibility of a controller must have the names/identity of all associated controllers - (CMQ2)

C.3 Each controller associated record of processing activities under the responsibility of (this) controller must have one or more contact details - (CMQ3)

C.4 Each controller associated with records of processing activities under the responsibility of (this) controller must have the identity and one or more contact details of the DPO - (CMQ4)

C.5 Processing under the responsibility of a controller must have one or more purposes of processing associated with it - (CMQ3)

C.6 Processing under the responsibility of a controller must have one or more categories of data subjects associated with it - (CMQ4)

C.7 Processing under the responsibility of a controller must have one or more categories of personal data associated with it - (CMQ5)

C.8 Processing under the responsibility of a controller must explicitly state when data is shared - (CMQ6)

C.9 Processing under the responsibility of a controller where data is shared must have the category of recipients to whom the data is or will be disclosed - (CMQ7)

C.10 Processing under the responsibility of a controller where data is shared must have the identities of recipients to whom the data is or will be disclosed - (CMQ8)

C.11 Processing under the responsibility of a controller where the recipients to whom the data is or will be disclosed are based in a Third Country or International Organisation must explicitly specified it as such - (CMQ9)

C.12 Processing under the responsibility of a controller must specify the identity of the Third Country or International Organisation to whom the data is or will be disclosed - (CMQ10)

C.13 Processing under the responsibility of a controller where personal data is transferred to a third country or an international organisation must specify the safeguards present for the transfer - (CMQ10)

C.14 Processing under the responsibility of a controller must explicitly state when data is

stored - (CMQ11)

C.15 Processing under the responsibility of a controller where data is stored must specify the criteria for its erasure - (CMQ12)

C.16 Each category of data associated with a record of processing must have a time limit or a condition or an event specified for its erasure - (CMQ13)

C.17 The record of processing must have technical and organisational security measures associated with it - (CMQ14)

C.18 Every sharing of personal data must specify the purposes of sharing with the recipients - (CMQ15)

C.19 Every processing must have information on how its records are being maintained - (CMQ16)

C.20 Each record of processing activity under the responsibility of a processor must have the names/identity of all associated processors under its responsibility - (CMQ17)

C.21 Each processor associated with a record of processing activity under the responsibility of a processor must have one or more contact details - (CMQ17)

C.22 Each processor associated with a record of processing activity under the responsibility of a processor must have the identity and one or more contact details of the DPO - (CMQ17)

C.23 Each record of processing activity under the responsibility of a processor must have the names/identity of the controller(s) it is acting on behalf of - (CMQ18)

C.24 Each record of processing activity under the responsibility of a processor must have name and contact details of the controller's representative and DPO - (CMQ18)

C.25 Each record of processing carried out by a processor on behalf of a controller must specify categories of processing associated with it - (CMQ19)

C.26 Each record of processing under the responsibility of a processor where data is shared must have the identity of recipients to whom the data is or will be disclosed - (CMQ20)

C.27 The identities of the recipients to whom data is or will be disclosed must be specified - (CMQ21)

C.28 If the recipients to whom the data is or will be disclosed are based in a Third Country, or International Organisation, this must be specified as such - (CMQ22)

C.29 The identity of the Third Country or International Organisation to whom the data is or will be disclosed must be specified - (CMQ22)

C.30 The transfer of personal data to a third country or an international organisation must specify its safeguards - (CMQ23)

C.31 The record of processing must have technical and organisational security measures associated with it - (CMQ24)

C.32 Every sharing of personal data must specify the purposes of sharing with the recipients - (CMQ25)

C.33 Each processing of data must have an associated legal basis - (CMQ26)

C.34 Each purpose for processing of data must have an associated legal basis - (CMQ27)

C.35 Each personal data must have information about its source - (CMQ28)

C.36 Personal data collected from the data subject must be clearly specified as such - (CMQ29)

C.37 Every anonymisation of data must specify the techniques used for anonymisation - (CMQ31)

C.38 Where pseudo-anonymised data can be de-anonymised by the organisation, it must be clearly specified as such - (CMQ32)

C.39 Every consent must be associated with only one Data Subject - (CMQ35)

C.40 Every consent must have one or more categories or types of personal data associated with it - (CMQ36)

C.41 Every consent must have one or more purposes associated with it - (CMQ37)

C.42 Every consent must have one or more processing associated with it - (CMQ38)

C.43 Every consent must have one and only one state/status - (CMQ39)

C.44 Every consent must be associated with one or more Controllers - (CMQ40)

C.45 Consent is given by exactly one Person - (CMQ41)

C.46 Consent provided by delegation must be clearly specified as such - (CMQ42)

C.47 Consent provided by delegation must have a single chain of delegation - (CMQ42)

C.48 Delegate in a consent has to play one or more roles that are associated with the Data Subject - (CMQ43)

C.49 Every delegation must have information on how it was executed - (CMQ44)

C.50 A delegate must be authenticated to act on behalf of the data subject in a delegation - (CMQ45)

C.51 Every consent must have information on who it was provided to - (CMQ46)

C.52 An entity collecting consent on behalf of the Data Controller must have information on the relationship - (CMQ47)

C.53 Every given consent must have information on how it was obtained - (CMQ48)

C.54 Every consent must have some artefacts associated with how it was given/obtained - (CMQ49)

C.55 Every consent must have information on what choices were provided to the data subject - (CMQ50)

C.56 Every consent must have a statement or affirmative action indicating given consent - (CMQ51)

C.57 Every consent must have information on how the right to withdraw was communicated - (CMQ52)

C.58 Consent must not have more than one location it was provided at - (CMQ53)

C.59 Consent must not have more than one medium it was provided in - (CMQ54)

C.60 Every consent must have a timestamp indicating when it was given/obtained - (CMQ55)

C.61 Every purpose or processing associated with Third Party must have information on the role played by the Third Party - (CMQ58)

C.62 If data is being stored, it must have information on how long it will be stored for - (CMQ60)

C.63 Every storage of data must have information on its storage location - (CMQ61)

C.64 Processing of personal data which is of automated nature must be clearly indicated as such - (CMQ62)

C.65 Every processing of data involving transfer to a third country or international organisation must have the identity of the third country or international organisation specified - (CMQ64)

C.66 Every personal data belonging to a special category must be clearly indicated as such - (CMQ65)

C.67 Every personal data must have information on one or more identifiers that link it to a particular data subject - (CMQ66)

C.68 A data subject who is not of legal age to provide their own consent must be clearly indicated as such - (CMQ67)

C.69 There must be information on the relevant laws that determine the legal age of consent - (CMQ68)

C.70 Every consent must be associated with only one Data Subject - (CMQ70)

C.71 Every consent must have one or more categories or types of personal data associated with it - (CMQ71)

C.72 Every consent must have one or more purposes associated with it - (CMQ72)

C.73 Every consent must have one or more processing associated with it - (CMQ73)

C.74 Every consent must have one and only one state/status - (CMQ74)

C.75 Every consent must be associated with one or more Controllers - (CMQ75)

C.76 Every change in the state/status of consent must be attributed to one or more agents - (CMQ76)

C.77 Consent changed by delegation must be clearly specified as such - (CMQ77)

C.78 Consent provided by delegation must have a single chain of delegation - (CMQ77)

C.79 Delegate in a consent has to play one or more roles that are associated with the Data Subject - (CMQ78)

C.80 Every delegation must have information on how it was executed - (CMQ79)

C.81 A delegate must be authenticated to act on behalf of the data subject in a delegation - (CMQ80)

C.82 Every change in the state/status of consent must have information on how it was changed - (CMQ81)

C.83 Every change in state/status of consent must have some artefacts associated with how it was changed - (CMQ82)

C.84 Every change to consent by the Data Subject must have a statement or affirmative action indicating the change - (CMQ83)

C.85 Consent must not have more than one location it was changed at - (CMQ84)

C.86 Consent must not have more than one medium it was changed in - (CMQ85)

C.87 Every consent must have a timestamp indicating when it was changed - (CMQ86)

C.88 Information about how the right to be informed was implemented must be specified - (CMQ88)

C.89 Information part of right to be informed must be concise - (CMQ88)

C.90 Information part of right to be informed must be transparent - (CMQ88)

C.91 Information part of right to be informed must be intelligible - (CMQ88)

C.92 Information part of right to be informed must be easily accessible - (CMQ88)

C.93 Information part of right to be informed must use clear and plain language - (CMQ88)

C.94 When the information relevant to the right to be informed was provided to the data subject must be specified - (CMQ89)

C.95 Information relevant to the right to be informed was provided to the data subject must be provided at most within one month of obtaining the data - (CMQ89)

C.96 If information relevant to the right to be informed is to be communicated to the data subject, it must be provided at most during the first communication with the data

subject - (CMQ89)

C.97 If data is to be disclosed to someone else, information relevant to the right to be informed must be communicated to the data subject at latest when the data is disclosed - (CMQ89)

C.98 Information part of right to be informed must contain name and contact details of the controller's representative - (CMQ90)

C.99 Information part of right to be informed must contain name and contact details of the DPO - (CMQ91)

C.100 Information part of right to be informed must contain purposes for processing - (CMQ92)

C.101 Information part of right to be informed must contain legal basis for processing - (CMQ93)

C.102 Information part of right to be informed must specify processing whose legal basis is legitimate interest - (CMQ94)

C.103 Where personal data is not obtained from the data subject, information part of right to be informed must specify categories of personal data obtained - (CMQ95)

C.104 Where personal data is not obtained from the data subject, information part of right to be informed must specify the sources of such data - (CMQ96)

C.105 Where personal data is shared, information part of right to be informed must specify the identity of the recipients or categories of recipients - (CMQ97)

C.106 Where personal data is transferred to a third country or international organisation, information part of right to be informed must specify the identity of the third country or international organisation - (CMQ98)

C.107 Where personal data is stored, information part of right to be informed must specify the retention period - (CMQ99)

C.108 Information part of right to be informed must specify the available rights - (CMQ100)

C.109 Where consent is obtained from the data subject, information part of right to be informed must specify the right to withdraw consent - (CMQ101)

C.110 Information part of right to be informed must specify right to lodge a complaint with the supervisory authority - (CMQ102)

C.111 Information part of right to be informed must specify where personal data is obtained under statutory or contractual obligation, - (CMQ103)

C.112 Information part of right to be informed must specify where personal data is obtained from the data subject under statutory or contractual obligation - (CMQ104)

C.113 Information part of right to be informed must specify if automated-decision making, including profiling is being used - (CMQ105)

C.114 Every record of a data breach must have a timestamp indicating when it occurred - (CMQ106)

C.115 Every record of a data breach must have a timestamp indicating when the controller became aware of the breach - (CMQ107)

C.116 Every data breach must be notified to the supervisory authority - (CMQ108)

C.117 Every record of a data breach must have a timestamp indicating when it was notified to the supervisory authority - (CMQ109)

C.118 Every record of a data breach must specify how the notification was provided to the supervisory authority - (CMQ110)

C.119 Every record of a data breach must specify the identities of the supervisory authorities it was communicated to - (CMQ110)

C.120 Every record of a data breach must specify if it is likely to result in a high risk to the rights and freedoms of the natural persons whose data is associated with it - (CMQ111)

C.121 Every record of a data breach must specify the process used to determine if it is likely to result in a high risk to the rights and freedoms of the natural persons whose data is associated with it - (CMQ111)

C.122 Every data breach must specify the data subjects affected by the data breach - (CMQ112)

C.123 Every data breach must be notified to the data subjects whose personal data was associated with the breach - (CMQ113)

C.124 Every record of a data breach must have a timestamp indicating when it was notified to the data subjects - (CMQ114)

C.125 Every record of a data breach must specify how the notification was provided to the data subjects - (CMQ115)

C.126 Every notification of a data breach to the data subject must provide information about the data breach - (CMQ116)

C.127 Every notification of a data breach to the data subject must provide information about mitigating potential effects of the data breach - (CMQ117)

C.128 Every record of a data breach must specify the data involved in the breach - (CMQ118)

C.129 Every record of a data breach must specify the technical measures for the protection of data involved in the data breach - (CMQ119)

C.130 Every record of a data breach must specify the steps taken to prevent or mitigate the effects of the data breach - (CMQ120)

SUMMARY

Through this chapter, an analyses of information associated with GDPR and its compliance was presented. Section 4.1 presented a model of information interoperability between entities as defined by requirements for GDPR compliance. The model provides an analyses of information requirements and information flows between different entities, and categorisation of information requirements as provenance records, consent information, compliance documentation, data processing agreements, and use of seals/certifications. Its analyses also provides a strong motivation towards adopting a common information model for representation of all activities associated with GDPR compliance.

Following this, Section 4.2 presented 'compliance questions' which aim to retrieve information relevant in evaluation of GDPR compliance. The section also provides the methodology used to formulate questions from authoritative sources. The compliance questions are accompanied with identification of assumptions and constraints which are useful towards establishment of information requirements and its validation.

The next chapter presents ontologies developed for addressing research objective $RO3$, and which utilise compliance questions presented in this chapter as competency questions in their ontology engineering process.

This page intentionally left blank.

5 | Representing Information for GDPR Compliance using Ontologies

This chapter presents OWL2 ontologies developed to fulfil research objective $RO3$ defined in Section 1.2.2 regarding representation of information. The chapter first presents a more detailed description of the methodology in Section 1.3.3 regarding developing and evaluating ontologies based on summary presented earlier in Section 5.1. It then presents the ontologies of: (i) GDPRtEXT (Section 5.2) which provides a linked data representation of GDPR text and a glossary of GDPR compliance concepts, and which satisfies research objective $RO3(a)$ by providing an ontological representation of concepts and text of GDPR; (ii) GDPRov Section 5.3 which enables representing provenance of activities associated with personal data and consent in ex-ante and ex-post phases, and fulfils research objective $RO3(b)$; and (iii) GConsent (Section 5.4) which enables representing information regarding consent, and fulfils research objective $RO3(c)$.

Each ontology is presented with a summary of its motivation, engineering process, and dissemination. Ontologies are presented with their evaluation based on the extent to which they satisfy the competency questions used in their development and through comparison with analysed approaches in state of the art in Section 3.7.

In addition to these, Data Privacy Vocabulary (DPV), initially presented in Section 1.4.6, is also included as an external contribution of the thesis based on work done by author of this thesis within the W3C Data Protection Vocabularies and Controls Community Group (DPVCG), and overlap of DPV with the research presented. Section 5.5 presents an overview of DPV and compares it with GDPRtEXT, GDPRov, and GConsent - and demonstrates their similarity in representing information while drawing attention to distinguishing features. The section also presents a comparison of DPV with state of the art as identified in Chapter 3.

5.1 Methodology for Ontology Engineering

5.1.1 Utilisation of Existing Ontology Engineering Methodologies

The creation of ontologies followed guidelines and methodologies deemed 'best practice' by semantic web community. In this, 'Ontology development 101: A guide to creating your first ontology' by Noy and McGuiness [35] was utilised as a guiding document for ontology creation. It provided steps for construction of an ontology with attention on avoiding bad design decisions and common pitfalls. It also suggested use of competency questions

to determine scope of an ontology and for evaluation after creation. The guide suggested Protégé[1] - a popular and widely adopted tool - for ontology development as it supports semantic reasoners to detect logical inconsistencies arising from asserted facts and axioms in ontology. Use of this guide provided a foundational basis for initiating the ontology development process and for using compliance questions from Section 4.2 as competency questions to identify concepts and relationships, testing for inconsistencies using Protégé, and iteratively building an ontology.

The development of ontologies followed a combination of NeOn methodology [34] and UPON Lite methodology [36]. NeOn provides a flexible workflow for ontology development through use of scenarios such a using a specification, reusing and re-engineering existing ontological and non-ontological resources, and utilisation of ontological design patterns. UPON Lite is a lightweight methodology for rapid ontology engineering that was used in combination with NeOn for iteratively developing ontologies in an agile fashion. UPON Lite consists of six steps: identification of domain terminology, construction of domain glossary, creating a taxonomy, predication as properties, meronymy for complex components, and conceptualisation into an ontology. The combination of NeOn and UPON Lite consisted of identifying development scenarios and specifying them as requirements using NeOn, then using UPON Lite to derive actionable tasks and implementing ontology creation.

The methodology used to develop ontologies presented in this chapter explicitly specifies competency questions used to derive its concepts based on compliance questions from previous chapter (Section 4.2). This approach enables tracing lineage of a concept to its role in compliance process and provides transparency in development process. To compare methodology used in this thesis with other methodologies used to develop ontologies within relevant approaches in SotA - some utilise legal experts which act as domain experts to validate developed ontologies and their interpretation - such as in research projects SPECIAL and MIREL (see Section 3.2). Such projects involve commercial partners who provide real-world use-cases and data to inform and evaluate developed research. Others - such as Ujcich et al. [50] - interpret GDPR as a set of requirements for compliance in their modelling of information. In either case, approaches within SotA do not provide competency questions that could be used to develop ontologies[2].

5.1.2 Ontology Quality

The quality of an ontology refers to quality of its design of concepts and relationships, and quality as a semantic dataset. While following a suitable ontology engineering methodology provides a structured ontology, it still needs to be inspected for quality in terms of ontology as well as for intended use-cases and scenarios. For this, existing publications [180], [181] list various methods of ontology quality detection, evaluation, and suggest solutions to fix identified problems.

[1] https://protege.stanford.edu/

[2] Deliverables of research project provide a description of how the concepts of their ontologies were developed from legal requirements, but such descriptions are argumentative and limited to the specified domain or use-case, and hence do not provide a concrete requirement that can be used to develop an ontology to answer the compliance questions.

OOPS![3] [38] is a useful tool for ontology evaluation which detects common pitfalls in design of concepts and relationships and provides a documented output which can be persisted for provenance of ontology development. Each pitfall detected by OOPS! is categorised along structural, functional, and usability-profiling dimensions. The tool also provides an indicative measure of importance regarding pitfalls in terms of critical, important, and minor levels. OOPS! was used for detecting catalogued common pitfalls in evaluation of developed ontologies. Identified pitfalls were corrected by changing underlying relationships to remove them.

Quality was also assessed and maintained by asserting sufficiency of developed ontology to represent and query information based on collected use-cases presented in Section 4.2.2.1. In this process, missing concepts and relationships were added to the ontology, while incorrect ones were removed or rectified.

5.1.3 Ontology Documentation

Ontology documentation was created by using WIDOCO[4] [37] - a tool which uses ontology metadata to create HTML documents listing its classes and properties. Ontology metadata consists of information regarding the ontology as well as its concepts and properties integrated into its serialisation as annotations. WIDOCO provides a document of suggested metadata indicating best practices for ontology documentation. It builds upon LODE[5] which is itself a popular ontology documentation service.

The output of WIDOCO is a HTML document along with various serialisations of ontology for content negotiation that can be published and used as an online resource. Additional information was manually added to HTML documentation to specify aims and methodologies used in development of ontologies as well as examples of use-cases and diagrams intended for human consumption. WIDOCO integrates OOPS! to detect pitfalls and documents the output. It also provides an interactive visualisation of the ontology using WebVOWL[6].

5.1.4 Dissemination

The ontologies were published on internet using a stable IRI through persistent identifiers on servers hosted by ADAPT Research Centre[7] and School of Computer Science & Statistics[8] within Trinity College Dublin. Initially, persistent identifiers were provided using purl[9] which later had issues regarding maintenance and frequent problems with URL resolution. The ontologies were then modified to utilise W3ID[10] persistent identifiers maintained by W3C Permanent Identifier Community Group[11]. The ontologies published in this manner followed best practices and principles related to use of Linked Open Data[12], Linked Open

[3] http://oops.linkeddata.es/
[4] https://dgarijo.github.io/Widoco/
[5] http://www.essepuntato.it/lode
[6] http://vowl.visualdataweb.org/webvowl.html
[7] https://adaptcentre.ie/
[8] https://scss.tcd.ie/
[9] https://purl.org/
[10] w3id.org/
[11] https://www.w3.org/community/perma-id/
[12] https://www.w3.org/TR/ld-bp/

Vocabularies[13], and FAIR[14] principles.

Each ontology was added to Linked Open Vocabularies[15] (LOV) - a community listing that catalogues vocabularies in semantic web community. Each ontology was published in Zenodo[16] which provides open repositories and assigns a unique DOI to repositories. The ontology and its resources were also added to public hosting repositories such as GitHub[17] and an instance of OpenGogs[18] hosted on institution servers. Each ontology was published under an open and permissive license (CC-by-4.0[19]) to promote its use and adoption.

5.1.5 Evaluation

Evaluation was carried out by analysing sufficiency of each ontology to provide concepts for representing information for answering competency questions. This was carried out in an iterative manner where each iteration consisted of developing the ontology, evaluating it, and utilising results of evaluation as feedback to identify areas of improvement such as missing concepts and relationships or incorrect assumptions.

The ontology was also evaluated against common pitfalls using OOPS! as described earlier regarding ontology quality. The OOPS! ontology report is published along with ontology documentation, and can be manually generated by using OOPS! online service. Documentation and publishing standards were evaluated by assessing whether ontologies met existing criteria advocated by the community (such as 5-star principle for linked data [20] and FAIR principles). Finally, each ontology was published and presented in a peer-reviewed venue and publication, with more information about publications provided in the respective ontology's section.

Evaluation of work as a research contribution was carried out based on whether it satisfied its research objectives motivating its development and whether it provided novel contributions compared to existing approaches within state of the art. The details of this are presented in evaluation sections of each ontology.

Summary of Methodology

Based on above description of ontology engineering processes, the methodology used for ontology engineering and development is summarised through as:

1. **Identification of aims, objectives, scope:** The first step was to identify aim and objectives of information to be represented, followed by deciding on scope regarding relation to GDPR compliance. For ontologies presented in this chapter, aims and objectives are listed in Section 1.2.
2. **Identify and analyse relevant information:** Using identified scope, relevant information was gathered from various sources including authoritative, community, and publications - and analysed to identify terms of importance and requirements regard-

[13] https://dgarijo.github.io/Widoco/doc/bestPractices/index-en.html
[14] Findability, Accessibility, Interoperability, and Reusability (FAIR) https://doi.org/10.1038%2Fsdata.2016.18
[15] https://lov.linkeddata.es/
[16] zenodo.org/
[17] github.com/
[18] opengogs.adaptcentre.ie/
[19] https://creativecommons.org/licenses/by/4.0/
[20] https://5stardata.info/en/

ing GDPR compliance. The information is presented as background of GDPR in Section 2.1 and analysed with regards to compliance in Chapter 4.

3. **Create use-cases and competency questions:** From the analysed information, different use-cases were identified to better understand application of information in compliance scenarios and requirements of different stakeholders in this process. This was done using information interoperability model presented in Section 4.1. The analysed information was used to create compliance questions, as presented in Section 4.2, which identify relevant information for evaluation of compliance. These compliance questions were utilised as competency questions in development and evaluation of ontologies.
4. **Identify concepts and relationships:** Relevant concepts and relationships were identified to express information required to answer compliance questions in identified use-cases. This was an iterative and cyclic process where identified concepts and relationships were re-purposed to better suit some design pattern or compliance requirements.
5. **Create Ontology:** The identified concepts and relationships were formalised as an ontology in OWL2 using the Protégé ontology development environment. In this process, a semantic reasoner (i.e. Pellet[21] and HermiT[22]) was used to identify logical inconsistencies in ontology. Minor inconsistencies were fixed by changing appropriate relationships between concepts, while major inconsistencies required evaluation of information identified in step 4. Development of ontology utilised best practices advocated by semantic web community in terms of ontology metadata, documentation, design patterns, publication, and dissemination.
6. **Evaluate:** The ontology was evaluated for sufficiency towards representing information for answering competency questions. The use of a semantic reasoner detected logical inconsistencies in expressed facts and axioms, while OOPS! provided detection of common pitfalls and bad design patterns. The quality of metadata and documentation was evaluated in terms of sufficiency based on community guidelines. Where an ontology was published and/or presented as a resource or as part of a peer-reviewed publication, resulting comments and feedback were used to identify areas of improvement. Citations of ontologies and associated publications were used to identify criticisms (if provided) and to compare them with work presented in citing publication.
7. **Dissemination:** The ontology and its documentation were published online with a persistent identifier as a FAIR resource with an open and permissive license. This included publication of ontology, datasets, and code in a public repository accompanied by human-readable documentation about its creation and utilisation.
8. **Progressive iterations:** Within a single iteration of development, an ontology was created and evaluated by following steps 2 to 6. Multiple iterations consisted of repeating these steps as in an development cycle to progressively improve the ontology by adding new concepts or removing existing undesirable ones. Previous versions of ontology were retained with their documentation for provenance where possible to indicate milestones in its development.

[21] https://github.com/stardog-union/pellet
[22] http://www.hermit-reasoner.com/

5.2 GDPRtEXT - Linked Open Dataset of GDPR text & Glossary of Concepts

This section describes the GDPRtEXT ontology and dataset which provides a linked data version of text of GDPR and a SKOS glossary of concepts associated with its compliance. The section presents the motivation and creation of GDPRtEXT, its publication, dissemination, and comparison with relevant approaches in state of the art. The latest iteration of GDPRtEXT (v0.6) is available online[23] with its documentation and code repository[24].

5.2.1 Motivation

GDPR as a legislation consists of text which is structured into 173 Recitals, 99 Articles (further grouped into Chapters and Sections), and 21 Citations. Each Article may have one or more Paragraphs which itself may have one or more Sub-Paragraphs. As per norms used in legislations, each individual clause - whether an article, paragraph, or sub-paragraph - is identified with an alphanumeric number as provided. These are commonly referenced in textual notation as identifiers, for example *Article 8 Paragraph 2 Sub-Paragraph c* can be referred to as: *A8(2-c)*, *A(8-2c)*, *A8-2(c)*, *Art.8 2(c)*, *Art-8-2-c*. As there is no standard or accepted commonality in specifying such references, and because such notations are intended for human readability and interpretation - a strict set or specification of notations does not exist. This presents difficulty when representing such information in machine-readable formats.

The EU Publications Office currently publishes legislation metadata at document level which provides information about GDPR as a legislation using ELI ontology and standard [39] but does not specify granular information about its contents - such as its articles. The EU Publications Office has indicated its intention to provide such granular metadata in future (see footnote in Section 3.7.2). Currently, concepts arising from legislations as well as those used in context of GDPR compliance have no standardised reference to provide commonality between two representations in different use-cases.

Within the larger scope of legal compliance, information is always associated with clauses and concepts of a law - in this case GDPR. Amongst approaches part of SotA presented in Chapter 3, only two approaches consider association of information with GDPR within scope of their work as presented in Section 3.7.2. Other approaches, where they reference concepts and clauses of GDPR, do so in an ad-hoc manner using textual notations such as "Article 4-11". As the analysis in Section 3.7.2 points out, the two approaches modelling clauses of GDPR have three drawbacks - (a) the representation is incompatible with ELI ontology, (b) none provide a glossary of terms relevant for compliance, and more importantly (c) neither resource can be reused as it is not published in an open and accessible manner. Addressing this gap is required to fulfil research objectives related to linking of information with GDPR.

With this motivation research objective $RO3(a)$ was established in Section 1.2.2 and is fulfilled by GDPRtEXT - which provides an OWL2 ontology for granular representation of GDPR text and a dataset created using this ontology for a linked data representation of GDPR where each clause has a unique IRI. GDPRtEXT also provides a SKOS glossary

[23] https://w3id.org/GDPRtEXT/
[24] https://github.com/coolharsh55/GDPRtEXT/

of terms associated with compliance and links them with their definition and relevance to clauses within GDPR using the developed ontology. It thus enables linking of information with specific clauses and concepts of GDPR.

5.2.2 Ontology Engineering and Creation of Resource

Following the methodology described in Section 5.1, development of competency questions was based on understanding and analysis of how legal articles are referenced in text in relation to compliance. Competency questions presented here are categorised based on whether they concern structure of GDPR or representation of concepts associated with compliance. These competency questions investigate structure and concepts of GDPR as a document and thereby differ from compliance questions presented in Chapter 4 which are concerned with investigating information for compliance. The competency questions for GDPRtEXT are outlined below with identified requirements:

5.2.2.1 Structure of GDPR text

`CQ.1` How many Recitals are there within GDPR?
`CQ.2` How many Chapters are there within GDPR?
`CQ.3` How many Sections are there within GDPR?
`CQ.4` How many Articles are there within GDPR?
`CQ.5` How many Paragraphs are there within GDPR?
`CQ.6` How many Sub-paragraphs are there within GDPR?
`CQ.7` How many References or Citations are there within GDPR?
`CQ.8` Article 4 belongs to which Chapter? (generalised to which *Chapter* does *Article X* belong?)
`CQ.9` Which clause contains the definition of 'personal data'? (generalised to definition of *concept X*)
`CQ.10` What is the structural hierarchy of the document?
`CQ.11` What are the 'Principles' defined in GDPR? (generalised to *conceptA* with types *conceptB*, e.g. 'Accountability' as a type of 'Principle')
`CQ.12` Which articles, paragraphs, and sub-paragraphs are relevant to the validity of given consent? (generalised to relevant to *concept X*)
`CQ.13` How to associate information regarding given consent to relevant clauses in the GDPR? (generalised to association information regarding *concept X*)
`CQ.14` How to associate information regarding compliance to a specific article of the GDPR?

Based on these, following requirements were identified with regards to extending the existing ELI ontology:

- Structure of text must be specified with granularity and a hierarchy of Document, Chapter, Section, Article, Paragraph, Sub-Paragraph along with Recitals and Citations.
- Relations between clauses must be specified e.g. Paragraph belongs to an Article.
- Relations must be transitive e.g. Paragraph in an Article must also be in the Article's Chapter.
- Each individual clause must have a unique IRI to enable linking of information to it.

5.2.2.2 Concepts associated with GDPR compliance

CQ.15 What type of data does the GDPR define?
CQ.16 What types of consent does the GDPR define?
CQ.17 What are the different entities referred to within GDPR?
CQ.18 Which activities are associated with processing of personal data?
CQ.19 Which activities are associated with consent?
CQ.20 What are the conditions or criteria associated which affect sensitivity of processing?
CQ.21 What activities are relevant to a data breach?
CQ.22 Which activities are relevant regarding compliance?
CQ.23 What are the principles defined in GDPR?
CQ.24 What are the rights provided by the GDPR?
CQ.25 Which criteria does the GDPR mention for right to data portability?
CQ.26 Which criteria does the GDPR mention for right to be informed?
CQ.27 What are the obligations mentioned within GDPR?
CQ.28 What are the obligations of the Controller?
CQ.29 What are the obligations of the Processor?
CQ.30 What are the obligations of a DPO?
CQ.31 What are the lawful basis for processing of personal data specified in the GDPR?
CQ.32 What are the conditions for valid consent under GDPR?
CQ.33 Which obligations are mentioned in relation to data collection?
CQ.34 Which obligations are mentioned in relation to obtaining consent?
CQ.35 Which obligations are mentioned in relation to retaining personal data?
CQ.36 Which obligations are mentioned in relation to security of personal data?
CQ.37 What concepts are defined regarding seals and certifications?

Based on these, following requirements were identified with regards to representing concepts as a glossary using W3C SKOS standard:

- The glossary should express concepts in a hierarchy of relation as associated with compliance. This hierarchy is based on which additional concepts are relevant to the given concept. For example, all principles are referred to when referring to the concept of 'Principle', and - activities and actions associated with compliance are referred to when using the concept of 'Compliance'.
- The glossary should reference concepts with their definitions within the clauses of the GDPR.
- The glossary should indicate relevant concepts within GDPR for a given concept in the context of compliance.
- The glossary should provide concepts regarding:
  - types of data
  - types of consent
  - types of entities
  - types of activities associated with - consent, data, processing, data breaches
  - actions associated with compliance
  - principles defined in the GDPR
  - rights provided by the GDPR
  - obligations mentioned in the GDPR

- conditions required for valid consent
- conditions associated with seals and certifications

5.2.2.3 Extending ELI

The suitability of extending existing ELI ontology for representing hierarchy of clauses in GDPR was evaluated and found to be feasible based on existence of generic extendible concepts. Information models used by ELI - namely FORMEX[25] and Common Data Model[26] were also taken into consideration in formulating an extension mechanism for compatibility. Where the GDPR as a document contains metadata regarding its expression in multiple languages, the extension was modelled to be language agnostic with labels in English for the scope of this thesis. Therefore language specifications provided by FRBR model[27] were not modelled nor included as language labels. The FRBR functionality can be easily integrated in future by extending relevant GDPRtEXT concepts with language tags and expressions.

5.2.2.4 Creation of datasets

Three outputs were decided to be produced under GDPRtEXT based on requirements stated earlier - an OWL2 ontology for representing structure of GDPR text, a dataset of GDPR text using this ontology, and a SKOS glossary of concepts associated with compliance. The OWL2 ontology and SKOS glossary were combined within a single deliverable - namely GDPRtEXT ontology, and representation of GDPR text was provided as a RDF dataset with metadata defined using DCAT[28] and VoID[29] standards.

The task of extracting individual clauses and annotating their structure (e.g. chapter, article, paragraph) was automated using a JavaScript script[30]. Three datasets were produced and published through this process. The first provides description of canonical versions of official legislation, i.e. those published by EU Publications Office which specifies GDPR legislation in HTML, PDF, and XML formats. The second provides a copy of GDPR hosted on institution servers and provides identifiers for individual clauses in HTML, JSON, and plain-text documents. The third provides RDF serialisations of GDPR using GDPRtEXT ontology in RDF/XML, N3, Turtle, and JSON-LD.

The glossary published using SKOS utilised IRIs of individual clauses of GDPR in GDPRtEXT to indicate source, definitions, and related concepts. Definitions were declared using `rdfs:isDefinedBy` property, and a new property called `gdprtext:involves` was created to indicate associations between concepts.

5.2.2.5 Publication & Dissemination

The ontology and dataset are exposed through a SPARQL endpoint[31] on a triple-store hosted on institution servers. Pubby[32] was used to provide a front-end for browsing a RDF seriali-

[25] https://op.europa.eu/en/web/eu-vocabularies/formex//
[26] https://op.europa.eu/en/web/eu-vocabularies/model/-/resource/dataset/cdm
[27] https://www.ifla.org/publications/functional-requirements-for-bibliographic-records
[28] https://www.w3.org/TR/vocab-dcat/
[29] https://www.w3.org/TR/void/
[30] https://github.com/coolharsh55/GDPRtEXT/blob/master/scripts/parse_gdpr.js
[31] https://w3id.org/GDPRtEXT/sparql
[32] http://wifo5-03.informatik.uni-mannheim.de/pubby/

sation of GDPRtEXT as shown in Figure 5.1. The dataset was provided under the CC-by-4.0 license to provide resources in an open and reusable manner and was published in Irish Open Data Portal[33] which rated it as a 5-star dataset indicating highest quality in following linked open data principles.

Figure 5.1: Article 12(3) in GDPRtEXT as RDF displayed using Pubby [72]

5.2.3 Resource Description & Application

An visual overview of concepts within GDPRtEXT is presented in Figure 5.2 and Figure 5.3.

5.2.3.1 Concepts for description structure of text

GDPRtEXT extends European Legislation Identifier (ELI) [39] ontology published by European Publications Office with granular concepts to represent individual clauses within GDPR. ELI provides the class `LegalResource` to indicate a legislative document and its sub-class `LegalSubResource` to indicate a component or part of that resource. GDPRtEXT extends `LegalSubResource` with sub-classes `Chapter`, `Section`, `Article`, `Point` (indicating Paragraph), `SubPoint` (indicating Sub-Paragraph), `Recital`, and `Citation`. ELI provides properties `has_part` and its inverse `is_part_of` to indicate connections between two legal resources, which GDPRtEXT extends using sub-properties to indicate hierarchical relations between chapters, sections, articles, points, and sub-points.

5.2.3.2 Concepts about Data

GDPR mentions different types of data which determine applicable obligations and requirements of compliance. GDPRtEXT provides `Data` as a top-level concept to indicate abstract term of 'data'. GDPR primarily focuses on personal data as defined in Article 4(1) - represented in GDPRtEXT as `PersonalData`, with special categories of personal data

[33] https://data.gov.ie/dataset/gdprtext

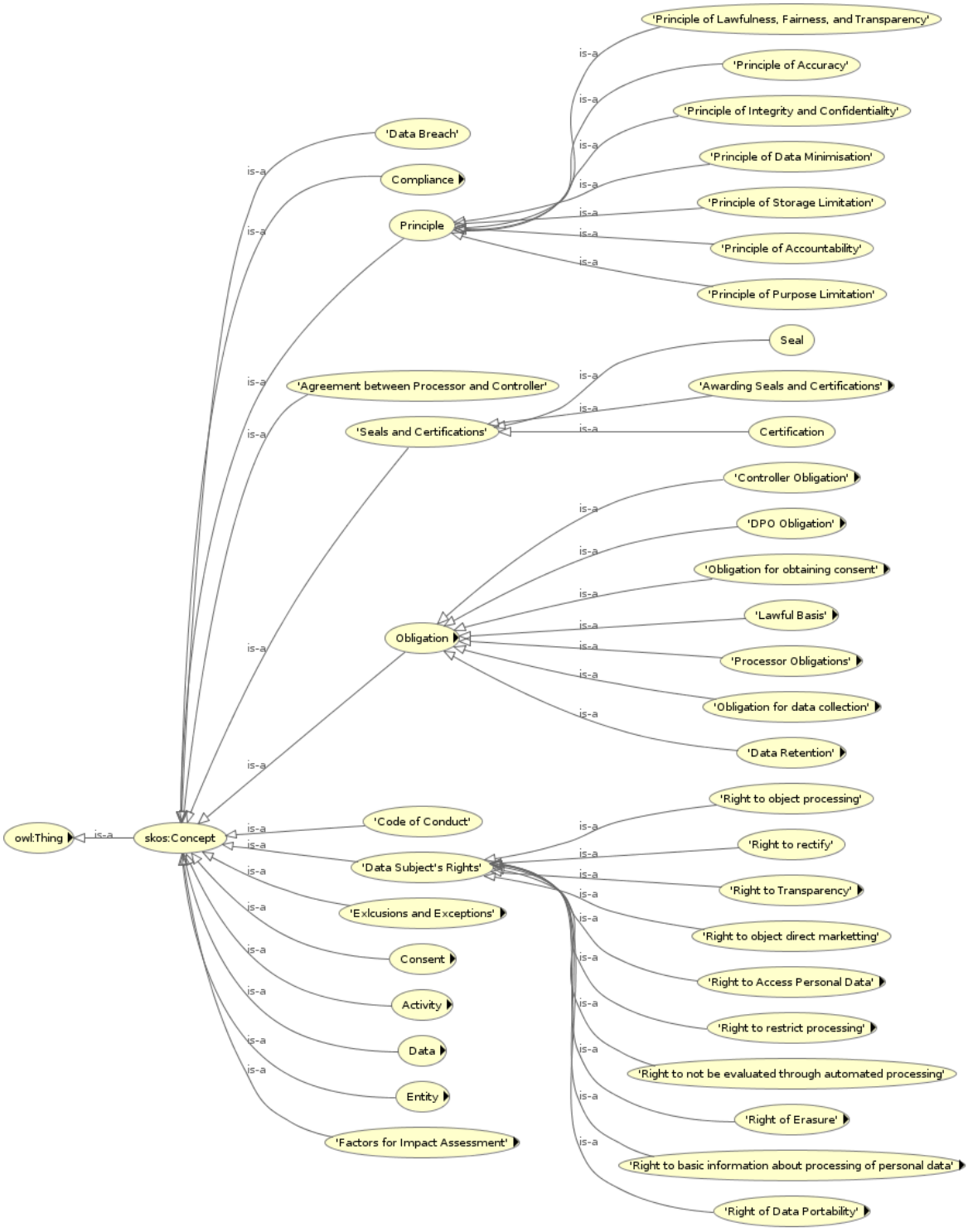


Figure 5.2: Visual overview of concepts in GDPRtEXT - part (a) [72]

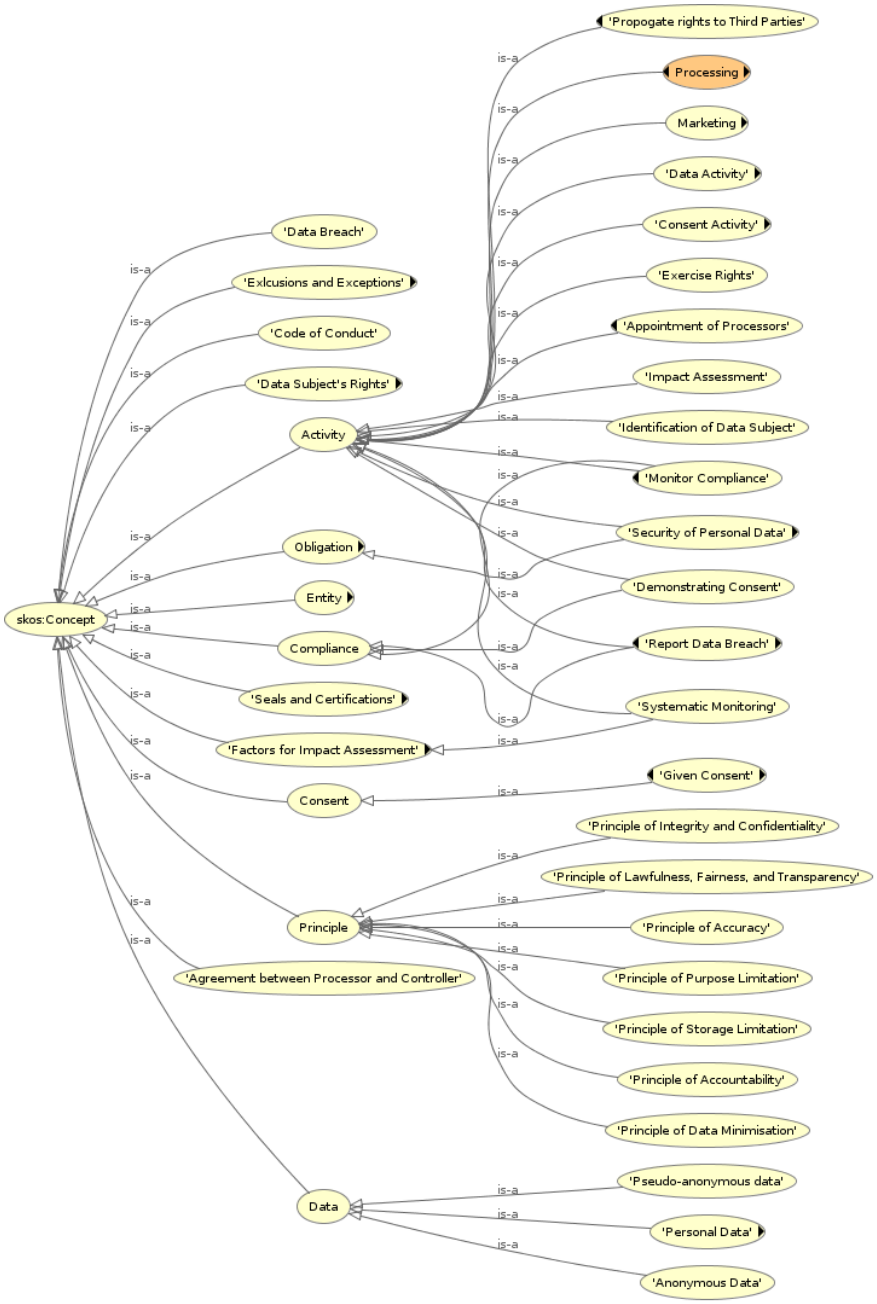


Figure 5.3: Visual overview of concepts in GDPRtEXT - part (b) [72]

defined in Article 9(1) requiring additional obligations for processing and handling being represented by `SpecialCategoryPersonalData`. Types of special categories mentioned include criminal data, genetic data, health data, and racial data - which are defined as sub-classes in GDPRtEXT. GDPR also mentions data in context of anonymisation and pseudo-anonymisation processes - represented in GDPRtEXT as `AnonymousData` and `PseudoAnonymousData`.

5.2.3.3 Concepts about Consent

The top-level concept of 'consent' is represented by `Consent` in GDPRtEXT with its definitions based in Articles 4(11), 6(1) and Recitals 32, 40. It is sub-classed as `GivenConsent` - which is a legal basis and therefore is also a sub-class of `LegalBasis`. `GivenConsent` is further sub-classed to indicate 'valid consent' which carries obligations of ensuring consent is valid and meets requirements of GDPR - and is therefore also defined as sub-class of `ObligationForObtainingConsent`. Obligations regarding conditions of valid consent are represented by sub-classing the `ValidConsent` for indicating - freely given, informed, specific, voluntary, and opt-in.

5.2.3.4 Concepts about Entities

`Entity` represents an 'entity' which could be an individual, institution, company, corporation, partnership, or government agency - to name a few. It is sub-classed to indicate entities specifically mentioned in GDPR: Data Subject, Controller, Processor, Sub-Processor, Data Protection Officer (DPO), and Data Protection Authority (DPA). Additionally, relevant concepts associated with entities are also defined: Representative of Controller, Representative of Processor, Certification Body, and Regulatory Authority.

5.2.3.5 Concepts about Activities

'Activity' refers to some process or action mentioned, referred, implied, or defined by requirements of GDPR compliance. To represent these, GDPRtEXT defines activities regarding consent and personal data processing, as well as other activities related to functioning of GDPR - such as reporting data breach and demonstrating consent. The top-level concept 'Activity' represents abstraction of all activities. 'ConsentActivity' and 'DataActivity' represent specialised activities involving consent and personal data respectively.

Consent activities defined within GDPRtEXT consist of obtaining consent and withdrawing consent. Data activities include use, archival, collection, cross-border transfer, erasure, copying, rectifying, sharing, and storage of personal data. In these, the activity associated with usage of personal data is equivalent to its common and synonymous usage with term 'processing'. Activities for indicating context of processing include - automated processing, automated decision making with significant effects, confirming or matching datasets, large scale processing, processing affected or vulnerable individuals, processing sensitive data, processing using untested technologies, and unlawful processing.

GDPRtEXT also provides activities associated with reporting of data breach, which includes obligations and actions such as - report data breach, maintain record of breach, notify data subject of breach, report breach to controller (for processors), and report breach to DPA

within 72 hours. Other activities provided are - security of personal data, appointment of processors, demonstrating consent, exercise rights, identification of data subject, impact assessment, marketing, direct marketing, monitor compliance, propagate rights to third parties, and systematic monitoring.

5.2.3.6 Concepts about Compliance

Concepts associated with compliance are provided to indicate actions or terms used in process of maintaining, documenting, evaluating, and demonstrating compliance. The top-level concept `Compliance` represents an abstract notion of compliance. Other terms derived from this include - Demonstration of Consent, Monitor Compliance, and Report Data Breach.

5.2.3.7 Concepts about Principles

GDPRtEXT represents principles using top-level concept `Principle`, which is specialised to indicate principles associated with: Accountability; Accuracy; Data Minimisation; Integrity and Confidentiality; Lawfulness, Fairness, and Transparency; Purpose Limitation; and Storage Limitation.

5.2.3.8 Concepts about Rights

To represent rights, GDPRtEXT provides top-level concepts representing each individual right with further concepts associated each right represented as sub-classes. The right of data portability is represented by `RightOfDataPortability` with related concepts regarding: providing copy of personal data, commonly used data format, machine readable format, structured, and supporting reuse.

The right of erasure is represented by `RightOfErasure` with related concepts provided regarding obligation to erase data when consent is withdrawn, or when data is no longer needed for original purpose. The right to access personal data is represented by concept `RightToAccessPersonalData` with related concepts for indicating if and where controller is processing data, whether there is automated processing with significant effects on data subject, categories of data being processed, categories of recipients data is shared with, existence of rights, information about processing, source of data, storage period, and ensuring no charges are levied for provision of rights.

`RightToBasicInformationAboutProcessing` represents right to basic information about processing and is accompanied with its related concept regarding information about third parties. The concept `RightToRestrictProcessing` represents right to restrict processing, and is accompanied with conditions such as - accuracy is contested, data no longer needed for original purpose, and processing is unlawful. The right to transparency is represented by `RightToTransparency` with related concepts regarding conditions of concise, easily accessible, intelligible, and transparent. Other represented rights include: right to not be evaluated through automated processing, right to object to direct marketing, right to object to processing, and right of rectification.

5.2.3.9 Concepts about Obligations

GDPRtEXT defines concepts regarding obligations of controllers, processors, DPOs, consent, and compliant processing of personal data based on a legal basis. Obligations of controllers are represented by `ControllerObligation` with related concepts provided regarding appointment of processors, accountability, controller responsibility, co-operation with DPA, data protection by design and default, data security, liability of joint controller(s), maintaining records of processing activities, privacy by design, propagate rights to third parties, and reporting data breach.

Rights of processors is represented by `ProcessorObligation` with related concepts for appointing sub-processors, assisting in complying with rights, compliance with controller's instructions, co-operating with DPA, data security, imposing confidentiality on personnel, informing controller of conflict with law, maintaining records of processing activities, only acting on documented instructions, propagating rights to third parties, providing controller with information for compliance, reporting data breach to controller, restrictions on cross-border transfers, and to return or destroy personal data at end of term.

The concept `DPOObligation` represents obligations of a DPO which include the monitoring of compliance represented by `MonitoringCompliance` . The obligations related to lawful basis for processing are represented by `LawfulBasisForProcessing` along with related concepts for contract with data subject, exempted by national law, employment law, given consent, historic, statistical, or scientific purposes, legal claims, legal obligation, legitimate interest, made public by data subject, medical or diagnostics use, not for profit organisation, public interest, purpose of new processing, and vital interest.

Obligations regarding valid consent are represented by `ValidConsent` with related concepts provided to indicate consent should be freely given, informed, specific, voluntary, and opt-in. Obligations for obtaining consent are represented by `ObligationForObtainingConsent` and include concepts for information about third parties, indicating consent can be withdrawn easily, and conditions regarding information provided for obtaining consent such as - it should be clear, providing explanation of processing, should not be from silence or inactivity, should be demonstrable, should be distinguishable from other matters, and that it should produce valid consent.

Obligations for data collection are represented by `ObligationForDataCollection`, which is accompanied with related concepts for indicating accurate collection, specification of explicit purpose, ensuring legitimate purpose, ensuring it is not further processed than original purpose, and ensuring it is limited to specified purpose. Obligations for retention of personal data are represented by `ObligationForRetentionOfPersonalData` and include related concepts about retention of personal data, ensuring it is adequate for processing, ensuring it is identifiable for required processing, obligation to kept it up to date, ensuring it is limited for processing, obligation to rectify inaccuracies, and ensuring it is relevant for processing. The concept `ObligationForSecurityOfPersonalData` represents obligations regarding security of personal data with related concepts provided regarding accidental loss, damage, destruction, and unlawful processing.

5.2.3.10 Concepts about Seals and Certifications

GDPRtEXT provides concepts of `Seal` and `Certification` for representing seals and certifications as provided by GDPR to assist with maintenance and demonstration of compliance. The conditions of these are represented by `ConditionsForSealsAndCertifications`, which is further expanded to represent conditions for seal/certification such as having a maximum validity of 3 years and having a voluntary system of accreditation.

5.2.3.11 Example Use-Case: Compliance Reporting

This example use-case, taken from documentation of GDPRtEXT [72], shows how references to GDPR can aid in creation of reports which document information regarding compliance.

Consider a system for creation of compliance reports that stores information related to obligations it addresses from GDPR. It uses the EARL[34] vocabulary for expressing results of conformance checks within the report. GDPRtEXT is used to link resources in EARL reports with articles and points within GDPR and to express and define concepts related to compliance in a suitable and comprehensible manner. Through this, information about compliance checks is linked and associated with specific articles of GDPR.

EARL provides a standardised vocabulary to describe specific resources and relationships that are relevant to test reporting. The core construct of EARL is an `Assertion`, which describes context and outcome of an individual test execution. It uses following concepts (copied verbatim from EARL specification):

- `Assertor` - This can include information about who or what ran the test. For example human evaluators, automated accessibility checkers, or combinations of these.
- `Test Subject` - This can include web content (such as web pages, videos, applets, etc.), software (such as authoring tools, user agents, etc.), or other things being tested.
- `Test Criterion` - What are we evaluating test subject against? This could be a specification, a set of guidelines, a test from a test suite, or some other testable statement.
- `Test Result` - What was the outcome of test? A test result could also include contextual information such as error messages or relevant locations.

Taking the example of Right to Data Portability, the EARL report in Listing 1 represents compliance checks for conditions associated with linked articles in GDPR (Article 20). The compliance system has a module `_system_dataportability` that checks software that handles provision of copy of personal data `_org_dataportability` through test case `_test_provide_data_copy` and generates a report showing the test has passed through `_result_pass`.

To gather related resources together, a SPARQL query (simplified) would focus on link between `TestCase` and its result using `earl:validity`, as shown in Listing 2. These tests can be further combined into test suites to group compliance checks related to each article or a particular concept and structure documentation around this form of logical grouping of concepts. In this manner, use of GDPRtEXT to link tests and results with documentation enables automation of information retrieval and management. A similar use-case of GDPRtEXT in linking constraints and their outcomes with GDPR is demonstrated in Chapter 6.

[34] https://www.w3.org/TR/EARL10-Schema/

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix earl: <http://www.w3.org/ns/earl#> .
@prefix dct:  <http://purl.org/dc/terms/> .
@prefix gdprtext: <http://purl.org/adaptcentre/resources/GDPRtEXT#> .
@prefix ex: <http://example.com/phd-thesis#> .

ex:org_dataportability
    a    earl:TestSubject, earl:Software ;
    dct:description "System that handles data portability requests"@en ;
    dct:title "Data Portability Handler"@en .

ex:system_dataportability
    a    earl:Assertor ;
    dct:description "Module checking data portability obligations"@en ;
    dct:hasVersion "1.4" ;
    dct:title "DataPortability Module"@en ;
    earl:asserts [ a earl:Assertion ;
      rdf:subject ex:org_dataportability ;
      rdf:predicate ex:result_pass ;
      rdf:object ex:test_provide_data_copy  ] .

ex:result_pass
    a    earl:ResultProperty ;
    earl:date "2018-01-01" ;
    earl:validity earl:Pass ;
    earl:confidence earl:High .

ex:test_provide_data_copy
    a    earl:TestCase ;
    earl:testMode earl:automatic ;
    dct:title "Test provision of data copy"@en ;
    dct:description "Tests data portability"@en ;
    dct:subject gdprtext:article20 .
```

Listing 1: Use of GDPRtEXT to link tests with GDPR Articles in EARL report

```
SELECT ?gdpr ?result ?confidence ?mode WHERE {
    ?assertor a earl:Assertor .
    ?assertor earl:asserts ?assertion .

    ?testcase rdf:predicate ?assertion .
    ?testcase a earl:TestCase .
    ?testcase dct:subject ?gdpr .
    ?testcase ear:testMode ?mode .

    ?testresult rdf:object ?assertion .
    ?testresult a earl:ResultProperty .
    ?testresult earl:validity ?result .
    ?testresult earl:confidence ?confidence .
}

| gdpr          | result     | confidence     | mode          |
-----------------------------------------------------------------
| article16     | pass       | low            | automatic     |
| article17     | pass       | high           | automatic     |
| article18     | fail       | high           | manual        |
| article19     | pass       | high           | automatic     |
```

Listing 2: SPARQL query and results showing retrieved GDPR test results by article

5.2.3.12 Example Use-Case: Mapping between DPD and GDPR obligations

The second application of GDPRtEXT, taken from its publication [72], demonstrates linking of obligations between GDPR and its predecessor - Data Protection Directive (DPD. Given that DPD was adopted in 1995, and was superseded by GDPR in 2016, there are a large number of solutions and approaches regarding compliance with DPD that already exist and are used in practice. By linking obligations between DPD and GDPR it is possible to investigate reuse of these existing solutions for GDPR compliance. To that end, a mapping from DPD obligations to GDPR obligations containing annotations that describe nature of changes is constructed by linking articles of DPD and GDPR.

To model annotations as a RDF resource, a linked data version of DPD was created similar to GDPRtEXT by assigning URIs for every clause in legislation. This enabled referring to each individual clause in DPD and linking it with relevant clauses in GDPR. The annotations (available online[35]) consist of references from a clause in DPD to its corresponding clause in GDPR with an expression of change between the two. The nature of change is represented by values: same - indicating no change; reduced - indicating reduction of obligation; slightly changed - indicating minor change; completely changed - indicating major change; and extended - indicating addition of obligations.

Its example demonstration consisted of using XACML[36] rules for controlling access to data and modelled after DPD obligations. For each link between DPD and GDPR obligations, a record was created indicating whether the corresponding XACML rules for DPD compliance needed to be changed to be applicable for GDPR. The notation `N/A` was used to denote cases where no XACML rules existed for a particular DPD obligation and where corresponding obligations in GDPR had changed or had additional requirements.

```
@prefix gdpr: https://w3id.org/GDPRtEXT/gdpr# .
@prefix dpd: https://w3id.org/GDPRtEXT/dpd# .
@prefix rdfs: http://www.w3.org/2000/01/rdf-schema# .

dpd:mappingrule6
    a dpd:DPDToGDPR_Annotation ;
    dpd:hasChange dpd:ChangeExtended ;
    dpd:hasXACMLChange dpd:XACMLNoChange ;
    dpd:resourceInDPD dpd:Article7 - a ;
    dpd:resourceInGDPR gdpr:Article6-1-a ;
    rdfs:comment "added consent given to ..." .
```

Listing 3: Example annotation of associating existing DPD compliance XACML rules with requirements of GDPR

The class `DPDToGDPR_Annotation` represents annotations between DPD and GDPR, with an example instance depicted in Listing 3. The property `resourceInDPD` is used to refer to a specific clause within DPD using its IRI. Similarly, the property `resourceInGDPR` is used to refer to a corresponding clause in GDPR. The nature of change is defined using property `hasChange` whose value is an instance of class `ChangeInObligation`, with instances defined for *Extended*, *Same*, *Reduced*, *CompletelyChanged*, and *SlightlyChanged*. Similarly, a change in XACML rules is defined using class `ChangeInXACMLRule` with instances

[35] https://openscience.adaptcentre.ie/projects/GDPRtEXT/dpd_mapping.html
[36] https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=xacml

Yes, *No*, and *N/A*.

5.2.4 Evaluation

In terms of ontology assessment, the methodology outlined in Section 5.1 provides criterion for evaluation of ontology quality and documentation. GDPRtEXT fulfils these based on using OOPS! tool[37] to identify and rectify bad design patterns and by following best practices and community guidelines for ontology documentation. GDPRtEXT and the work described in this section was published [72] at Extended Semantic Web Conference (ESWC) - Resource Track. The publication described creation of resource, summarised its contents, and described mapping of DPD obligations with GDPR using a linked data approach and XACML to denote which obligations from DPD could be re-used towards GDPR compliance. ESWC is a premier and top-tier conference within semantic web domain, and has a rigorous review process with an open review policy. The acceptance of GDPRtEXT in this venue demonstrates its value as a semantic web resource.

To date, the publication has received 19 citations from peer-reviewed publications (excluding self-citations) on Google Scholar[38]. In addition to these, a 5 star rating given to GDPRtEXT as a dataset in Irish open data portal indicates its adherence to linked data principles. A survey of legal approaches within state of the art [43] undertaken by MIREL project analysed GDPRtEXT amongst other legal ontologies and found that GDPRtEXT is singular in its use of ELI and provision of GDPR as a glossary of concepts - a finding shared with the analyses of SotA in Section 3.7.

5.2.4.1 Fulfilment of Competency Questions

The assessment of GDPRtEXT consists of evaluating the extent to which it answers competency questions outlined in Section 5.2.2. For this, Table 5.1 shows concepts and relationships of GDPRtEXT relevant towards answering competency questions.

Table 5.1: Concepts in GDPRtEXT for answering competency questions

CQ	Concepts/Relationships
CQ1-7	*Recital, Chapter, Section, Article, Point, SubPoint, Reference, Citation*
CQ8	*isPartOfChapter*
CQ9,11	*rdfs:isDefinedBy [Article, Point, SubPoint]*
CQ10	*:_ hasPart/isPartOf :_*
CQ12	*:GivenConsent rdfs:seeAlso [Article, Point, SubPoint]*
CQ13,14	*GivenConsent/Compliance :involves [Article, Point, SubPoint]*
CQ15	*Data, PersonalData, SensitivePersonalData, CriminalData, GeneticData, HealthData, RacialData, AnonymousData, PseudoAnonymousData*
CQ16	*Consent, GivenConsent, WithdrawnConsent*
CQ17	*Entity, DataSubject, Controller, JointController, Processor, SubProcessor, DPO, DPA, ControllerRepresentative, ProcessorRepresentative, CertificationBody, RegulatoryAuthority*
CQ18,19	*DataActivity, ConsentActivity*

(Cont'd on following page)

[37] OOPS! results published with ontology documentation. The results can also be independently obtained using the OOPS! online service.

[38] https://scholar.google.com/scholar?cites=2776106745007214232

Concepts in GDPRtEXT for answering competency questions (cont'd)

CQ	Concepts/Relationships
CQ20	*Processing, AutomatedProcessing, AutomatedDecisionMakingWithSignificantEffect, ConfirmingOrMatchingDatasets, LargeScaleProcessing, ProcessingAffectedOrVulnerableIndividuals, ProcessingSensitiveData, ProcessingUsingUntestedTechnologies, Unlawful Processing*
CQ21	*ReportDataBreach, MaintainRecordOfBreach, NotifyDataSubjectOfBreach, ReportBreachToController, ReportBreachToDPAWithin72Hours*
CQ22	*Compliance, Demonstration, ConsentMonitor, Compliance, ReportDataBreach*
CQ23	*Principle, Accountability, Accuracy, DataMinimisation, IntegrityAndConfidentiality, LawfulnessFairnessAndTransparency, PurposeLimitation, StorageLimitation*
CQ24	*Rights, RightOfDataPortability, RightOfErasure, RightToAccessPersonalData, RightToTransparency, RightToBasicInformationAboutProcessing, RightToNotBeEvaluatedThroughAutomatedProcessing, RightToObjectForDirectMarketting, RightToObjectToProcessing, RightToRectify, RightToRestrictProcessing*
CQ25	*RightOfDataPortability, ProvideCopyOfPersonalData, ShouldBeCommonlyUsedFormat, ShouldBeMachineReadable, ShouldBeStructured, ShouldSupportReuse*
CQ26	*RightToBasicInformationAboutProcessing, InformationAboutThirdParties*
CQ27,28	*Obligation, ControllerObligation, AppointmentOfProcessors, Accountability, ControllerResponsibility, CooperateWithDPA, DataProtectionByDesignAndDefault, DataSecurityLiabilityOfJointControllers, MaintainRecordsOfProcessingActivities, PrivacyByDesign, PropogateRightsToThirdParties, ReportDataBreach*
CQ29	*ProcessorObligation, AppointingSubprocessors, AssistInComplyingWithRights, ComplianceWithControllersInstructions, CooperateWithDpa, DataSecurity, ImposeConfidentialityOnPersonnel, InformControllerOfConflictWithLaw, MaintainRecordsOfProcessingActivities, OnlyActOnDocumentedInstructions, PropogateRightsToThirdParties, ProvideControllerWithInformationForCompliance, ReportDataBreachToController, RestrictionsOnCross-borderTransfers, ReturnOrDestroyPersonalDataAtEndTerm*
CQ30	*DPOObligation, MonitorCompliance*
CQ31	*LawfulBasisForProcessing, ContractWithDataSubject, ExemptedByNationalLaw, EmploymentLaw, GivenConsent, HistoricStatisticalOrScientificPurposes, LegalClaims, LegalObligation, LegitimateInterest, MadePublicByDataSubject, MedicalDiagnosticOrTreatement, NotForProfitOrg, PublicInterest, PurposeOfNewProcessing, VitalInterest*
CQ32	*ValidConsent, FreelyGivenConsentObligation, InformedConsentObligation, SpecificConsentObligation, VoluntaryOptInConsentObligation*
CQ33	*ObligationForDataCollection, AccurateCollection, ExplicitPurpose, LegitimatePurpose, NotFurtherProcessedThanOriginalPurpose, SpecifiedPurpose*
CQ34	*InformationAboutThirdParties, ConsentCanBeWithdrawnEasily, ClearExplanatinOfProcessing, NotFromSilenceOrInactivity, Demonstrable, DistinguishableFromOtherMatters, ValidConsent*
CQ35	*RetentionOfPersonalData, AdequateForProcessing, IdentifiableForRequiredProcessing, KeptUpToDate, LimitedForProcessing, RectifyInaccuracies, RelevantForProcessing*
CQ36	*SecurityofPersonalData, AccidentalLoss, Damage, Destruction, UnlawfulProcessing*
CQ37	*Seal, Certification*

The table demonstrates that GDPRtEXT provides concepts to answer all competency questions. GDPRtEXT thus meets requirements of representing and linking information with text and concepts of GDPR in a granular manner and fulfils $RO3(a)$.

5.2.4.2 Comparison with SotA

The SotA in representing text of GDPR in machine-readable formats presented in Section 3.7.2 compared three approaches: ELI [39], Agarwal et al [40], and PrOnto [41], [55]. Their comparison and analysis, summarised in Table 3.2, depicts relevance of each approach in representing the GDPR as a glossary of concepts, providing a permanent identifier for resources, modelling of GDPR's text, and whether resources are open and accessible. The conclusion drawn from these is the lack of an approach fulfilling all criteria along with a lack of open and reusable resources concerning GDPR. The additional resource of ELI+ mentioned in analysis shows intention of EU Publications Office to remedy this gap through an update to the ELI ontology at some time in future.

A comparison of GDPRtEXT with these approaches, depicted in Table 5.2, shows that GDPRtEXT provides a glossary of concepts, uses permanent identifiers, provides linked data version of text of GDPR, and is available under an open and permissive license (CC-BY-4.0). This matches the intended contributions of ELI+ (update to ELI) planned by EU Publications Office, and therefore enables GDPRtEXT to fill this gap in this time.

Table 5.2: Comparison of GDPRtEXT with SotA

Work	**GDPRtEXT**	ELI	ELI+	Agarwal et al	PrOnto
Vocabulary	ELI	OWL2	OWL2	RDFS	Akoma Ntoso
Granularity	Sub-Paragraph	Legislation	Sub-Paragraph	Paragraph	Sub-Paragraph
Glossary	✓	✗	✓	✗	✗
PID	✓	✓	✓	✗	✗
OA	✓	✓	✓	✗	✗
GDPR text	✓	✗	✓	✗	✓

A survey of legal ontologies by Leone et al. [43] includes GDPRtEXT as an ontology relevant for data protection. The survey also includes ELI and PrOnto within the scope of data protection ontologies - which provides external comparison between these and GDPRtEXT. The survey outlines the role of GDPRtEXT in acting as a glossary of concepts rather than a prescriptive set of norms and rules for specification of compliance - such as made available through PrOnto. In this role, GDPRtEXT is novel within state of the art given a lack of other similar resources.

Based on this, GDPRtEXT is argued to provide novel contribution to state of the art and addresses gaps associated with representation of concepts and GDPR text at a granular level, and whose open availability enables usage and adoption.

Summary

The GDPRtEXT resource represents the first major contribution of this thesis. It provides a linked data version of text of GDPR and a glossary of its concepts, fulfils research objective $RO3(a)$, and assists with research objective $RO5(b)$ - as outlined in Section 1.2. It enables representing each article or point within GDPR as a unique resource through IRIs defined using RDF and semantic web. GDPRtEXT thus enables machine-readable links to be es-

tablished between information and clauses of GDPR as well as concepts pertaining to its compliance.

The use of GDPRtEXT makes it possible to create approaches that automate generation and querying of information associated with GDPR - such as for compliance, management of business processes, or generation of privacy policies. The compatibility provided through extension of ELI ontology ensures alignment with official documents produced by European Publications Office. Finally, GDPRtEXT fills an important gap in the state of the art regarding machine-readable approaches for linking information with legal text. GDPRtEXT has been released as an open resource, has been published in Zenodo and Datahub, and has been incorporated into Ireland's open data portal as a 5-star linked open dataset.

5.3 GDPRov - Ontology for GDPR activities associated with Personal Data and Consent

This section describes the GDPRov ontology for representing activities in ex-ante and ex-post phases associated with processing of personal data and consent for GDPR compliance. GDPRov stands for GDPR Provenance - a reference to the requirement of maintaining provenance information of processes in both ex-ante and ex-post phases for demonstrating GDPR compliance. This section presents motivation, overview, dissemination, and evaluation of GDPRov ontology. It also presents comparisons with relevant approaches in state of the art.

The ontology satisfies the research objectives $RO3(b)$ presented in Section 1.2. It uses the compliance questions presented in Section 4.2 as competency questions to identify requirements and for evaluation. GDPRtEXT is used to define and associate the source of concepts within the text of GDPR. An earlier version (v0.4) of GDPRov was described in a peer-reviewed publication [66]. Subsequent revisions included addition of new concepts associated with real-world implementation and interpretation of GDPR compliance requirements (see Section 6.1) and for representing information about consent mechanisms on the internet (see Section 6.2). The latest version of GDPRov (v0.7) is available online[39] with its documentation and code repository[40].

5.3.1 Identification of requirements from competency questions

The compliance questions presented in Section 4.2 were selected based on relevance to information regarding activities and provided competency questions for deriving concepts and relationships regarding processes associated with personal data and consent based on GDPR compliance requirements. These concepts and relationships were collected, combined, and analysed to ensure their cohesion as an ontology and evaluated using compliance questions to ensure they satisfied requirements regarding GDPR compliance and documentation of associated processes. In this, ex-ante and ex-post representation provide repetition of some information as most processes have counterparts in both phases. The linking of information between phases enables them to be documented in a manner so as to demonstrate prior planning of processes to ensure compliance and later their execution which also needs to be documented for compliance. Therefore, while GDPR requirements and compliance

[39] http://w3id.org/GDPRov
[40] https://github.com/coolharsh55/GDPRov

questions do no explicitly mention or refer to ex-ante and ex-post phases for each activity, GDPRov implicitly considers each activity to have representations in both phases.

The sub-sections below present concepts to answer competency questions derived from compliance questions. This is followed by an analysis of discovered concepts in ex-ante and ex-post phases. The analysis is used to derive requirements for construction of GDPRov ontology, and is presented to describe motivations for the ontology design and implementation.

5.3.1.1 Actors and Agents involved in activities

- CMQ2 - Provides concept of *Controller* as an agent controlling processes and its representative *Data Protection Officer (DPO)*.
- CMQ17 - Describes *Processor* as an executor of processes and its representative *DPO*. In this relationship, *Controller* provides processes to be executed as instructions to *Processor* through a *Data Processing Agreement (DPA)*.
- CMQ35 - Describes *Data Subject* as an agent who is associated with provision of personal data, consent, and who is related to exercising of rights.

5.3.1.2 Details of processing

- CMQ3 and CMQ37 provide concept of *Purpose* which describes purpose of personal data processing. Each purpose can incorporate multiple processing operations, and each processing operation taking place can be associated with multiple purposes.
- CMQ4 describe necessity to specify data subject categories whose personal data is being processed.
- CMQ36 describes personal data, while CMQ5 describes categories of personal data being processed. CMQ34 specifies special categories of personal data as a sub-category of personal data that needs to be explicitly stated as being processed.
- CMQ38 defines processing of personal data as defined by Article 4-2 of GDPR. The GDPR definition of processing provides types of operations as specified by "*any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction;*".
- CMQ6 defines sharing of data as a type of processing. Additional information associated with sharing of data is provided by - CMQ7 and CMQ20 for categories of recipients; CMQ8,CMQ21 for identifies of recipients, CMQ9 and CMQ22 for location where data is being sharing to; CMQ10 and CMQ23 for safeguards associated with data transfer; CMQ15 and CMQ25 for purposes of sharing - which is the same concept as purpose of processing except applied for sharing of personal data.
- CMQ11 defines data storage, with additional concepts provided by CMQ12 for existence of time limits or conditions for erasure and CMQ13 for specification of time limits or conditions for erasure for categories of data.
- CMQ26 defines legal basis for justifying processing of personal data, and CMQ27 specifies legal basis associated with a particular purpose. Each purpose can have one or more legal basis associated with it.

5.3.1.3 Life-cycle of data

- CMQ28 and CMQ30 describe source of personal data which in turn implies an activity that collects data and specifies actors or agents as source of data.
- CMQ29 specifically refers to personal data collected from data subject.

5.3.1.4 Anonymisation

- CMQ31 specifies anonymisation of personal data with CMQ32 inquiring about different 'levels' of anonymisation which affect the application of obligations and requirements of compliance.
- Levels of anonymisation are specified based on their relevance to investigation of compliance, and are defined as data which is: completely anonymised, pseudo-anonymised, not anonymised. In this, data that is pseudo-anonymised can be considered and used as anonymous data when an organisation does not have additional information to de-anonymise it.
- Processing activities associated with anonymisation and de-anonymisation of personal data are defined to produce anonymised data.

5.3.1.5 Activities associated with Consent

- Regarding consent, CMQ48 inquires about activities associated with provision and collection of consent. This includes information about how consent is requested and collected, used within processes as a legal basis, and is archived for future demonstration of compliance.
- CMQ49 and CMQ50 inquire about artefacts associated with collection of consent for determination of consent validity under GDPR, which requires investigation of how choices for consent were offered. This also includes forms through which consent is provided or collected from data subjects. Artefacts such as forms or dialogues are associated with processes where consent choices are offered or requested and whose result is collection of consent or given consent.

5.3.1.6 Provision of Rights

- The rights associated with GDPR need processes to internally (in perspective of an organisation) handle their execution as well as for interactions with data subjects. Therefore, such processes need to be defined and documented for compliance purposes.
- For right to be informed, CMQ88 – CMQ105 provide competency questions regarding how the right is provided and how it is executed or implemented.
- This includes activities associated with provision of information to data subject, artefacts associated with information provision, inclusion of details such as controller and DPO, purposes, processing, legal basis, and personal data categories.
- It also includes information about sources of personal data (where not obtained directly from data subject), and whether legal basis used is legitimate interest.
- Regarding data sharing, information to be specified includes categories of recipients and their location.

- Right to be informed also includes provision of information regarding existence and application of rights.
- Information associated with right to be informed is common to other information documented in due course of processing of personal data, and therefore does not require a separate notation or representation of this information in order to execute the right. Existing information or concepts for representation of processing activities can be reused for specifying required information. However, activities associated with executing rights need to be defined to demonstrate existence of processes for handling rights.

5.3.1.7 Compliance procedures such as Reporting of Data Breach

- Reporting of data breach requires information about data breach to be maintained as specified by `CMQ106 - CMQ120`.
- This includes information about the data breach consisting of: timestamp of when breach occurred (`CMQ106`), timestamp of when controller became aware of it (`CMQ107`), timestamp and method of it being notified to supervisory authority (`CMQ108`).
- Information about contents of breach include information about its affected personal data and categories of data subject (`CMQ112`).
- Information also needs to be provided to supervisory authorities and in some cases to data subjects based on extent of breach (`CMQ113`). It therefore requires prior planning of processes for handling data breaches and sending information to data subjects along with any remedial measures (`CMQ116`).

5.3.1.8 Specifying requirements for ex-ante and ex-post phases

Process logs are a convenient and demonstrable form of information to store and document compliant processing of personal data. By verifying logs, it is possible to document, evaluate, and demonstrate that executed processes were compliant with requirements of GDPR. This constitutes as ex-post phase of compliance and consists of evaluating information after processing has been carried out. It also fulfils Article 30 of GDPR concerning processing records to be maintained. Along with ex-post records, it is also essential to demonstrate that executed processes were based on a preconceived plan or template that was ensured to be compliant before execution. Storing such plans is essential to demonstrate prior planning and maintenance of a compliant processing system. This constitutes as ex-ante phase of compliance, and consists of evaluating compliance on plans of processing yet to be carried out. This is necessary for Article 35 of GDPR concerning carrying out a DPIA.

Associating executed processes with their plans allows demonstration of compliance throughout the life-cycle of processes, i.e. from planning of processes to their eventual execution. It also enables documenting changes in plans and their effects on execution of processes - i.e. demonstrating that when a plan changes it also brings about corresponding changes in executed processes. In context of GDPR compliance, requirements of compliance require documentation, maintenance, and demonstration of processes across both ex-ante (planning) and ex-post (execution) phases. Ex-ante plans of processes are described as an organisational measure and their compliance is associated with ensuring processes meet legal requirements before they are actually carried out. In some instances, such as for a Data

Protection Impact Assessment (DPIA), existence of ex-ante information about processes is essential in evaluation of compliance.

While compliance questions provide a basis for identifying information to be modelled, requirements of expressing this information in ex-ante and ex-post phase require specification of their intended usage in planning and processing stages respectively which further determines whether compliance evaluation consists of verification of a plan or analysis of processing logs. The argument and motivation for representing processes in ex-ante and ex-post phases represents a design decision based on separating representation of information across phases of compliance rather than being a compliance requirement itself. Approaches within state of the art that also follow a similar representation of ex-ante and ex-post information include SPECIAL (Section 3.2) which uses PROV-O to log information in both phases, and MIREL (Figure 3.2) which uses a workflow model to represent a plan and its executions.

Information requirements for modelling information about activities is summarised through following points:

1. Represent process in ex-post phase as a log or record.
2. Represent process in ex-ante phase as plan or template.
3. Link ex-ante plan with its instantiations or executions in ex-post phase.
4. Track provenance of ex-ante plans i.e. changes in plans of processes.
5. Enable tracking changes in ex-post logs based on corresponding changes in ex-ante plans.
6. Associate information used/generated in activities as artefacts in both ex-ante and ex-post phases.
7. Associate actors/agents with processes.
8. Link processes based on:
   (a) dependency - where one process is dependant on another through use of generated artefact,
   (b) order of execution - where one process is or will be executed before or after another, and
   (c) composition - where one process is constituted by several sub-processes.

5.3.2 Extending PROV-O and P-Plan

Based on above stated requirements for representing activities or processes in ex-ante and ex-post phases, existing semantic web ontologies of PROV-O [47] and P-Plan [48] were extended with relevant GDPR concepts and relationships to create the GDPRov ontology. The necessity of this process and a brief overview of PROV-O and P-Plan ontologies is presented below along with the process of extending ontologies.

5.3.2.1 PROV - W3C standard for representing provenance information

Provenance is information about entities, activities, and people (or software) involved in producing data or a component which can be used to form an assessment about its quality, reliability, or trustworthiness. The PROV-O ontology [47] along with PROV family[41]

[41] https://www.w3.org/TR/2013/NOTE-prov-overview-20130430/

of schemas and documents is the W3C recommendation for representing provenance information since 30th April 2013 and has seen significant adoption by semantic web and industrial communities. It provides definitions for interchange of provenance information by representing entities and relations between them such as generated by, derived from, and attributions.

The core concepts of PROV-O are summarised in Figure 5.4 and consist of interactions between *Activities*, *Entities*, and *Agents*. An `Entity` in PROV-O is defined as being physical, digital, conceptual, or other kind of 'thing' with some fixed aspects. PROV-O defines an `Activity` as something that occurs over a period of time and acts upon or with entities; it may include consuming, processing, transforming, modifying, relocating, using, or generating entities.

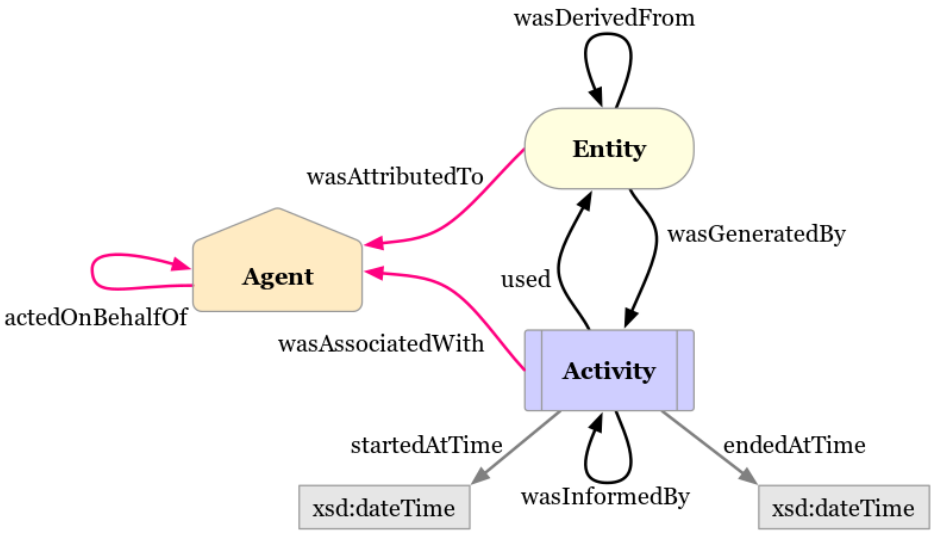


Figure 5.4: Overview of PROV-O model [47]

PROV-O is a generic and domain independent ontology for representing provenance information. In order for it to be applied in domain of GDPR compliance, it needs to incorporate relevant terminology and enable distinction between different types of activities and entities. Furthermore, PROV-O as a provenance ontology is intended to represent information about activities that have been executed in the past, and is therefore suitable to represent information only in the ex-post aspects.

While PROV-O does provide the concept of `Plan`[42] to represent ex-ante information, it does not provide further concepts or relationships to associate a plan with activities and entities[43]. In order to adopt PROV-O and use `Plan` for representing ex-ante information for GDPR compliance, it needs to be extended with additional concepts and relationships.

[42] PROV-O defines a *plan* as a set of actions or steps towards some goal. It clarifies on the lack of concepts relevant to plans as - "*There exist no prescriptive requirement on the nature of plans, their representation, the actions or steps they consist of, or their intended goals.*"

[43] PROV-O provides the concept of `Association` which assigns responsibility to an agent for an activity and indicates that the agent had a role in the activity, which can include a `Plan` associated using the `hasPlan` property.

5.3.2.2 P-Plan - extending PROV-O Plans as Workflows

P-Plan [48] extends the concept of `Plan` in PROV-O towards representing scientific workflows which enable creating a template of a 'step' and linking it to executions of activities. A `p-plan:Plan` is a subclass of `prov:Plan` and is composed of smaller activities or steps (`p-plan:Step`) that use and generate (as inputs or outputs) variables (`p-plan:Variable`). An overview of relationship between PROV-O and P-Plan is described in Figure 5.5. P-Plan enables representation of provenance information associated with both ex-ante and ex-post processes by representing them as scientific workflows. It also enables associating plans with their executions, thereby providing a link between ex-ante and ex-post provenance information.

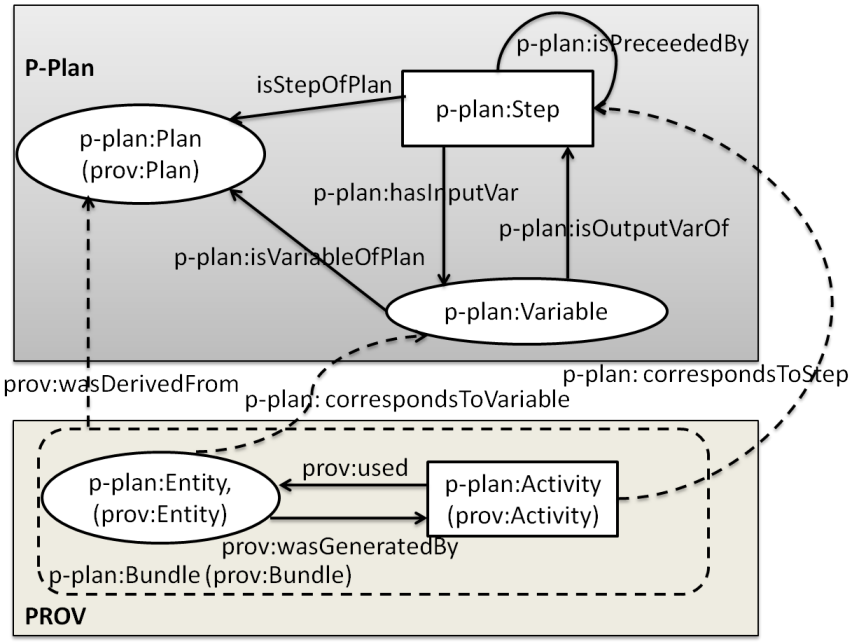


Figure 5.5: Overview of P-Plan model and its relationship with PROV-O [48]

A `p-plan:Plan` represents information of 'how' something should happen or a 'template' for executions. A `p-plan:Activity` is a subclass of `prov:Activity` and represents execution of process described in a `p-plan:Step`. A `p-plan:Entity` is a subclass of `prov:Entity` that corresponds to a `p-plan:Variable` in `p-plan:Plan`. Therefore, a `p-plan:Step` may describe the template including inputs and outputs which can then be instantiated into multiple instances of `p-plan:Activity` that can have distinct inputs to produce different outputs. As `p-plan:Plan` extends `prov:Plan`, which itself extends `prov:Entity`, it can be used to treat `p-plan:Plan` as an object whose provenance can be tracked using PROV-O or P-Plan. This makes it possible to express 'provenance of provenance' - thereby creating a record of how plans were formulated and executed over time.

5.3.2.3 Extending ontologies for GDPR

The PROV-O and P-Plan ontologies were extended to represent concepts and relationships of ex-ante and ex-post activities associated with personal data and consent based on requirements of GDPR compliance. The decision to extend PROV-O and P-Plan with GDPR concepts was made as both ontologies contain generic concepts associated with activities

and workflows which can be used for representing information about GDPR compliance, but doing so would be not be intuitive due to difference in terminology and structuring of information as expected for GDPR compliance.

Extending existing ontologies of PROV-O and P-Plan enables expressing a 'template' or 'plan' using `p-plan:Plan` describing ex-ante activities (as `p-plan:Step`) that can take place. This template can then be used to denote execution of activities in ex-post phase using `p-plan:Activity`. This provides a machine-readable and documented information model for both ex-ante and ex-post activities whose provenance itself can be expressed (using PROV-O and P-Plan) to record how they were created and how they change over time. This is beneficial in documenting state of a system at a given time through a set of activities that deal with consent and personal data, and can be helpful in determining changes to be made based on changes in processing of personal data over time.

The extended ontology derived from PROV-O and P-Plan incorporates concepts and relationships associated with GDPR in order to normalise terminology for representing information associated with GDPR compliance. Concepts and relationships are derived from competency questions and linked with relevant clauses within GDPR through using GDPRtEXT concepts using `rdfs:isDefinedBy` and `rdfs:seeAlso`. This provides documentation regarding origin of concepts and their use in representation of information associated with specific clauses of GDPR. It also provides a machine-readable link from the ontology to GDPR which can be used to compare, analyse, and align relevant ontologies.

The extension consists of sub-classing existing concepts in PROV-O and P-Plan to represent specific activities associated with GDPR compliance. The use of subclass mechanism preserves existing concepts and relationships of PROV-O and P-Plan to provide compatibility and reuse. This is particularly important for PROV-O as it is a W3C standard and therefore is more likely to have existing uses. The compatibility with PROV-O also enables information defined using the ontology to be bundled as an artefact to record its provenance and planning as a form of meta-documentation regarding planning and maintenance of compliance activities This is particularly useful to maintain periodic snapshots of organisational processes associated with compliance and provides opportunities to automate querying and validation of information within a use-case - as demonstrated in Chapter 6.

5.3.3 Ontology Description & Application

The ontology engineering is named GDPRov (GDPR Provenance Ontology) and is published online along with its documentation at https://w3id.org/GDPRov/ under an open and permissive CC-by-4.0 license. The ontology was created, documented, and published using the methodology presented in Section 5.1. The aim of GDPRov is to provide representations of ex-ante and ex-post activities regarding personal data and consent for GDPR compliance. It uses GDPRtEXT concepts to define origin and relevance of its concepts to GDPR.

5.3.3.1 Overview of GDPRov concepts

GDPRov extends concepts from PROV-O and P-Plan using the sub-class and sub-property mechanisms to represent activities associated with GDPR compliance, with a visual overview provided in Figure 5.6.

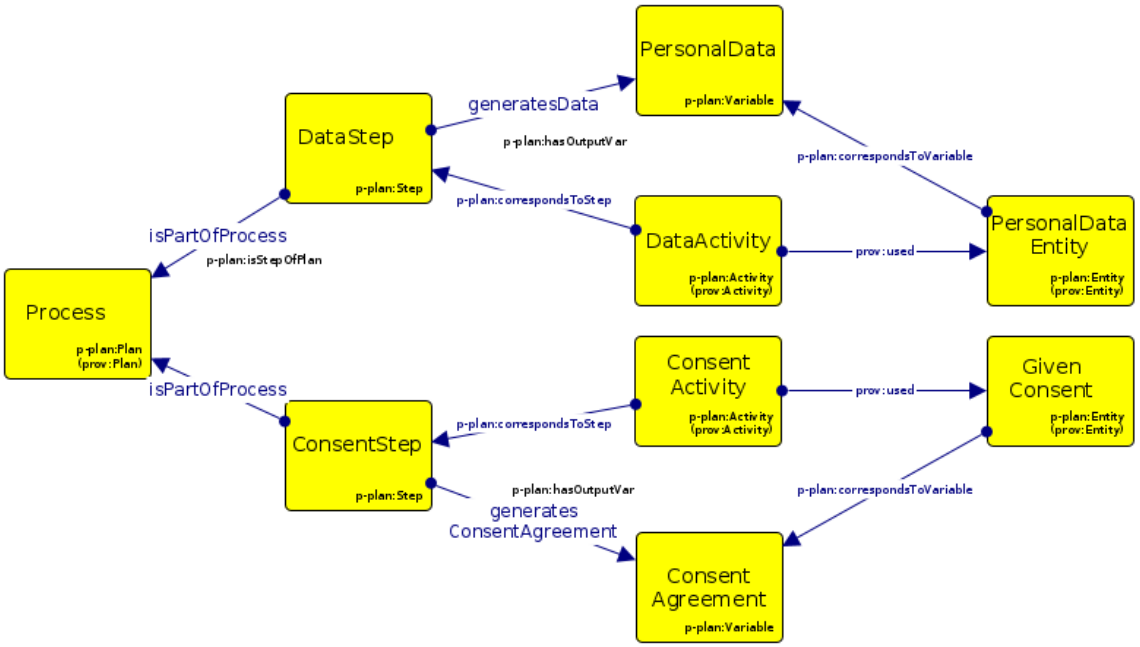


Figure 5.6: GDPRov concepts derived by extending PROV-O and P-Plan

GDPRov extends `p-plan:Plan` as `Process` to represent ex-ante plans of activities that will take place. The terminology is based on common use of term in expressions such as 'business processes' and 'compliance processes'. Each `Process` can contain steps (represented by `p-plan:Step`) to represent activities that interact with data and agents. To associate steps with a process, the property `p-plan:isStepOfPlan` is extended as `isPartOfProcess`. Another additional property - `refersToProcess` is also used to enable referring to a process without being a part of it. Similarly, to associate data (defined in P-Plan as `p-plan:Variable`) the properties `p-plan:hasInputVar` and `p-plan:isOutputVarOf` are extended for activities using inputs and producing outputs respectively.

Ex-post activities in P-Plan are represented by `p-plan:Activity`. Data interactions with these activities is represented by `p-plan:Entity` and the properties `prov:used` and `prov:wasGeneratedBy` are used to indicate inputs and outputs respectively. GDPRov defines steps to indicate automated execution and user interactions regarding collecting data from user (input) and providing data (output). To indicate a legal basis associated with a process or a step, the property `hasLegalBasis` is provided.

5.3.3.2 Depicting Data Life-cycle

Activities associated with life-cycle of personal data constitute of collecting, processing or using it, storing, sharing, deleting, transferring, transforming, anonymise, and rectifying data. GDPR defines several more categories of actions in Article 4-2 in its definition of 'processing'. GDPRov provides broad and abstract processes to represent data access, data archival, data erasure, and data rectification given the need to execute these using one or more steps. GDPRov provides representations of actions in ex-ante phase as `DataStep` which extends `p-plan:Step` and in ex-post phase as `DataActivity` which extends `p-plan:Activity`. These are further extended to distinguish between data collection, data

deletion, data sharing, data storage, data archival, data transfer, data transformation, data usage, and rectification of data. A visual overview of steps describing a data life-cycle using GDPRov is provided in Figure 5.7.

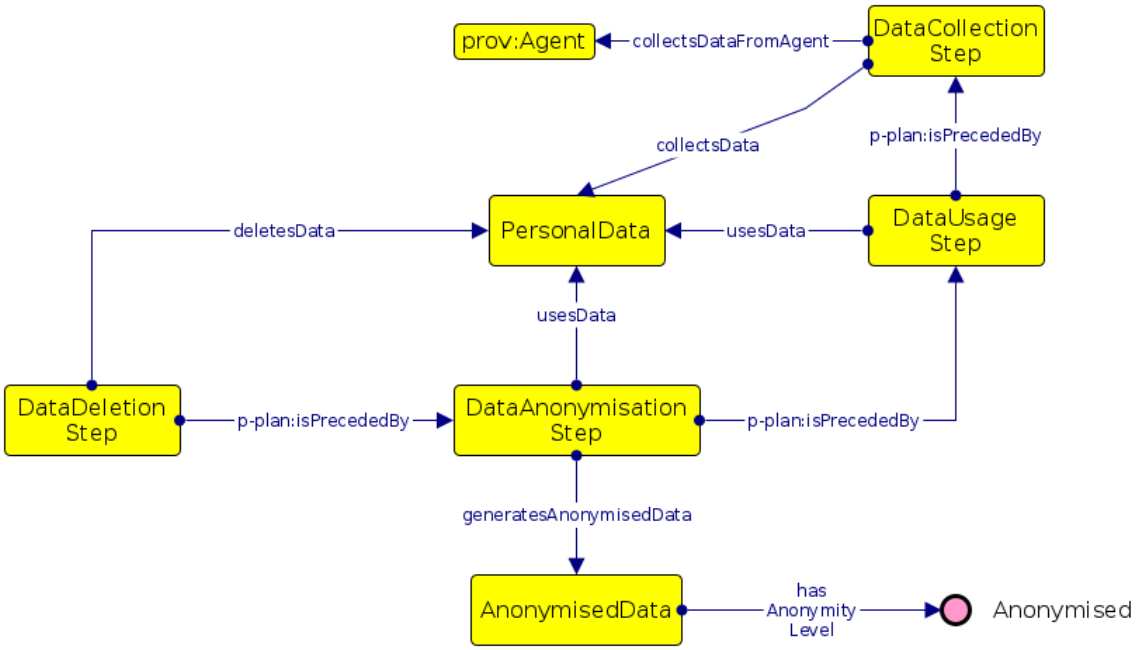


Figure 5.7: Example steps depicting data life-cycle using GDPRov

Anonymisation of data is defined as a sub-class of data transformation to indicate transformation of data that takes place when anonymising it. As GDPR obligations are based on level of anonymity and capability of de-anonymising it from an organisation's point of view, GDPRov provides the concept of `anonymisation level` to indicate state of anonymity of data. GDPRov defines four levels of anonymisation based on existing work in representing anonymous data [182], which constitute of data that is: (i) completely anonymised, (ii) completely de-anonymised, (iii)pseudo-anonymised, and (iv) pseudo-organisational-anonymised where an organisation does not have additional data required to de-anonymise it and can thus utilise it internally as if it were completely anonymous data. Sharing of data consists of interactions with actors or agents, which are represented by `prov:Agent` and associated with respective steps and activities using extended properties.

Personal data used within activities is represented by `PersonalData` which is sub-classed from `p-plan:Variable` for ex-ante representation and by `PersonalDataEntity` which is sub-classed from `prov:Entity` for ex-post representation. Further categorisation of personal data into anonymised, sensitive, and representing user identifier is provided through sub-classes.

5.3.3.3 Depicting Consent Life-cycle

Activities associated with the life-cycle of consent are represented in ex-ante phase by sub-classing `p-plan:Step` as `ConsentStep` and similarly in ex-post phase by sub-classing `p-plan:Activity` as `ConsentActivity`. These are further sub-classed to represent acquisition, archival, modification, and withdrawal of consent. Amongst these, withdrawal of

consent is defined as sub-class of modification since it modifies state of consent. A visual summary of steps in a consent life-cycle is provided in Figure 5.8.

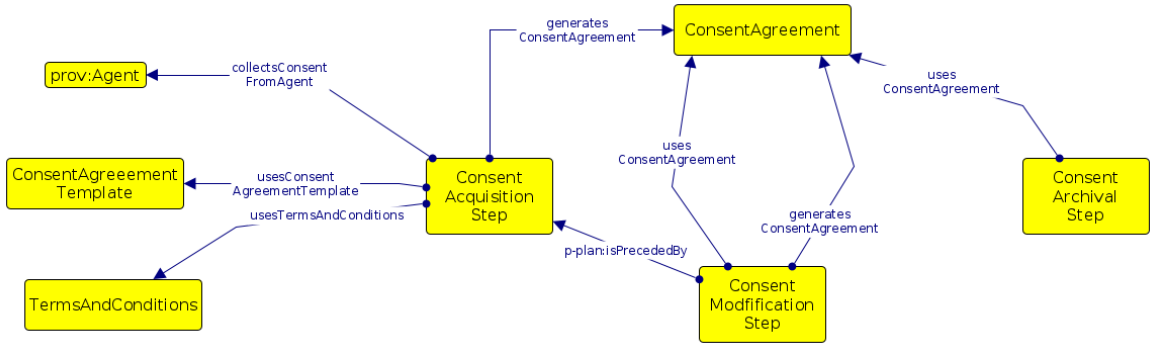


Figure 5.8: Consent life-cycle defined using GDPRov

Artefacts associated with consent and used in activities include choices or offer provided to request consent and subsequent consent given by an individual. To represent these in ex-ante phase, GDPRov provides concepts of `ConsentAgreementTemplate` to represent template offered to collect consent, `ConsentAgreement` to indicate given consent, and `TermsAndConditions` to indicate applicable policies or terms and conditions. Corresponding concepts in ex-post phase are `GivenConsentTemplate`, `GivenConsent`, and `TermsAndConditionsEntity`.

5.3.3.4 Depicting Compliance-related processes

In addition to representing activities associated with personal data and consent, GDPRov also provides representations for compliance-related processes. These include actions such as appointing processor (by a controller), carrying out an impact assessment, marketing and its special case of direct marketing, and monitoring compliance. Processes are also provided for handling data breaches which include notifying controller (by a processor), notifying data subject, and notifying data protection authority. The handling of rights required by GDPR is represented through sub-classes of `Process` for data portability, erasure, access personal data, basic info about processing, no automated processing, object to direct marketing, object processing, rectification, restrict processing, transparency, SAR (subject access request).

5.3.3.5 Example Use-Case: Querying anonymised sharing of data

The applicability and usefulness of GDPRov is demonstrated through its use for querying and validation of information for GDPR compliance in Chapter 6. A simplified example demonstrating a similar application through use of a SPARQL query was published along in a peer-reviewed publication [66]. It is presented in Listing 4 to demonstrate how GDPRov can assist in answering of compliance questions for GDPR.

The query uses GDPRov concepts to retrieve data being shared, specific steps that share it, anonymisation level of shared data, and steps used to anonymise it. The query is meant to retrieve information relevant in investigation of data being shared and its anonymity.

```
PREFIX gdprov: <https://w3id.org/GDPRov#>
SELECT ?data ?sharestep ?isAnonymised ?anonymisationStep
WHERE {
    ?data a gdprov:Data .
    ?sharestep a gdprov:DataSharingStep .
    ?sharestep gdprov:sharesData ?data.
    BIND (
        EXISTS { ?data a gdprov:AnonymisedData . }
        as ?isAnonymised ) .
    OPTIONAL {
        ?anonymisationStep
        gdprov:generatesAnonymisedData ?data .
    }
}
```

Listing 4: SPARQL query representing compliance question `G5` concerning legal basis for processing

5.3.3.6 Example Use-Case: Detecting changes in activities for updates to consent

As an use-case, consider the case where a data controller updates a plan of processing activities - such as when a purpose changes or a new processing operation is added to an existing purpose, and where legal basis for such processing is consent. In such cases, a data controller is required to evaluate whether updating an individual's consent is required based on changes between given consent and new purposes or processing activities. By storing plans of processing operations using GDPRov, it is possible to compare the old and new versions of a plan, detect changes, and identify whether corresponding updates to consent are needed.

An exploration of above as change detection was published in Managing the Evolution and Preservation of the Data Web workshop co-located with ESWC 2018 [46]. It described comparison of plans represented using P-Plan to identify changes based on the above obligation.

The change detection, visualised in Figure 5.9, is based on identifying differences between two plans in terms of steps and variables and whether they have been added, removed, or modified. In the figure, the change reflects removal of a step - which by itself does not require any changes to given consent since no new purposes have been added to an existing given consent. Using this approach, the detected change can be analysed - manually for complex and legal interpretations and automatically for simpler or simplified graphs - and used to identify whether corresponding changes are necessary based on compliance obligations.

5.3.4 Evaluation

The ontology assessment was based on the methodology outlined in Section 5.1 regarding criterion for ontology quality and documentation. The ontology was used in two applications developed to demonstrate use of SPARQL to query information (see Section 6.1) and SHACL to validate information for GDPR compliance (see Section 6.2). The experience demonstrates suitability of GDPRov in representing the required information, and led to addition of `ConsentAgreementTemplateBundle` (in v0.7) as a concept for convenience

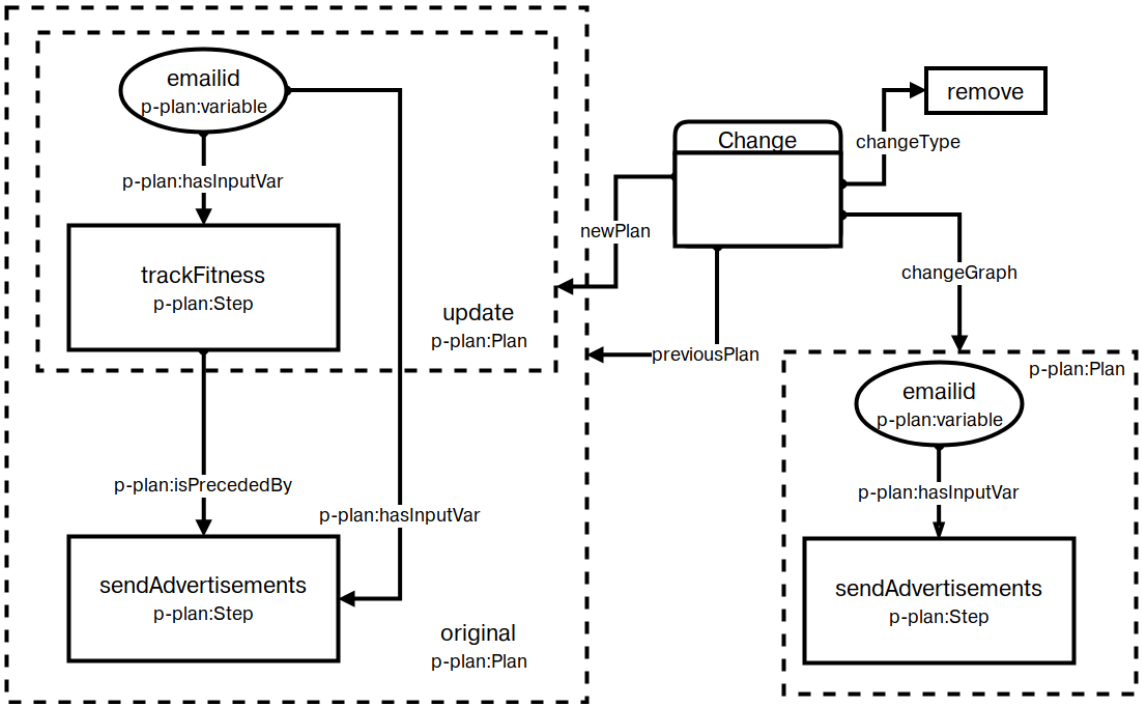


Figure 5.9: Modelling changes in workflows using P-Plan [46]

in representing 'bundled' consent requests and provisions based on consent workflows on a website where a single dialogue is used to collect consent involving multiple distinct purposes and third parties.

GDPRov was published [66] as a peer-reviewed publication in Workshop on Society, Privacy and the Semantic Web - Policy and Technology (PrivOn) co-located with the 16th International Semantic Web Conference (ISWC). The workshop provided reviews from domain experts in privacy, legal, and semantic web domains; with ISWC being a top-tier conference in semantic web domain. As of February 2020, this publication has received 18 citations to (excluding self-citations) on Google Scholar[44] of which 2 are deliverables of CitySPIN research project (see Section 3.2). A publication describing an approach for annotating DFDs (data flow diagrams) with information for analysing compliance [183] utilised GDPRov to represent personal data as an entity used in activities within its ontology for representing DFDs to abstract processing operations as data flows.

5.3.4.1 Fulfilment of Competency Questions

An assessment of the extent to which GDPRov satisfies competency questions by providing concepts and relationships is presented here as part of its evaluation. The competency questions, summarised in Section 5.3.1, were used to guide development of ontology and therefore are used to evaluate the extent to which the developed ontology meets requirements of representing this information. Table 5.3 lists concepts and properties for answering competency question (with *N/S* used to indicate not in scope).

[44] https://scholar.google.com/scholar?cites=2287149512924017207

Table 5.3: Concepts in GDPRov for answering competency questions

CQ	Class	Property	Phase
Actors and Agents involved in activities			
CMQ2	*Controller, ControllerRepresentative, DPO*		
CMQ17	*Processor, ProcessorRepresentative, DPO*		
CMQ35	*DataSubject*		
Details of processing			
CMQ3	*Process*	*refersToProcess*	*Ex-ante*
CMQ4	*DataSubject*		
CMQ5	*PersonalData, SensitiveData*	*usesData*	*Ex-ante*
	PersonalDataEntity, SensitiveDataEntity		*Ex-post*
CMQ6	*DataSharingStep*	*sharesData, sharesDataWith*	*Ex-ante*
	DataSharingActivity	*hasSharedDataWith*	*Ex-post*
CMQ7	*ThirdParty*	*sharesDataWithThirdParty*	*Ex-ante*
CMQ8	*ThirdParty*	*sharesData, sharesDataWith*	*Ex-ante*
CMQ9	*N/S*	*N/S*	
CMQ10	*N/S*	*N/S*	
CMQ11	*DataStorageStep*	*usesData, generatesData*	*Ex-ante*
	DataStorageActivity		*Ex-post*
CMQ12	*N/S*	*N/S*	
CMQ13	*N/S*	*N/S*	
CMQ26		*hasLegalBasis*	
Lifecycle of data			
CMQ28	*DataCollectionStep*	*collectsData*	*Ex-ante*
	DataCollectionActivity		*Ex-post*
CMQ29	*DataCollectionStep*	*collectsDataFromAgent*	*Ex-ante*
	DataCollectionActivity	*collectedDataFromAgent*	*Ex-post*
Anonymisation			
CMQ31	*DataAnonymisationStep, AnonymisedData*		*Ex-ante*
	DataAnonymisationActivity, AnonymisedDataEntity		*Ex-post*
CMQ32	*PersonalData, SensitiveData*	*hasAnonymityLevel*	*Ex-ante*
Activities associated with Consent			
CMQ48	*ConsentStep, ConsentAcquisitionStep, ConsentModificationStep, ConsentArchivalStep, ConsentAgreement, ConsentAgreementTemplate*	*usesConsentAgreement, generatesConsentAgreement*	*Ex-ante*
	ConsentActivity, AcquireConsentActivity, ArchiveConsentActivity, ModifyConsentActivity, GivenConsent, GivenConsentTemplate	*collectedConsentFromAgent*	*Ex-post*
CMQ49	*ConsentAgreementTemplate*		*Ex-ante*
	GivenConsentTemplate		*Ex-post*

Questions not in scope (marked as *N/S* either require clarity from authoritative sources regarding interpretation of information to provide a concrete design pattern or have multiple possible representations of which it cannot be determined which is more useful from a legal compliance point of view. Examples include location of recipients - which can be expressed either through an property/annotation associated with a data sharing activity or attached with a particular third party; and specifying time limits or duration or conditional events associated with data storage and deletion periods. These have been identified as

future work regarding further development of the ontology based on differing interpretations of representation, complexity of specifying values such as "EU membership" and "as long as required", and pending expert opinion of legal authorities on these issues through courts or executive decisions. These are considered minor issues regarding representation of information as they do not have a major impact on design and use of GDPRov. Approaches and ontologies in SotA provide alternative design patterns which can be used as non-authoritative approaches for short-term mitigation of this issue.

The presented evaluation demonstrates GDPRov satisfies requirements of answering competency questions regarding representation of activities and identifies those that are needed to be resolved as future work based on availability of legal opinion and decisions. GDPRov thus fulfils research objective $RO3(b)$ by providing representations of activities associated with personal data and consent in ex-ante and ex-post phases.

5.3.4.2 Comparison with SotA

The representation of process flows and activities associated with GDPR compliance in existing approaches was presented and analysed as part of state of the art in Section 3.7.3. The attributes for this analysis involved features or concepts that could be represented using the specified approach and basis for representation in existing vocabularies and standards. The analysis demonstrated existence of a variety of approaches that utilised existing standards of PROV-O and BPMN to model GDPR-specific information regarding process flows and activities in both ex-ante and ex-post phase. It found that approaches modelling both ex-post and ex-ante phases exist and utilise PROV-O as their basis for representation of information.

A comparison of GDPRov with SotA is provided in Table 5.4 using the same attributes used for analysis. The table lists features supported by each approach using a check mark (✓) with a blank indicating no information regarding the feature was found. Column headings corresponding with expression of information supported by an approach, and use following abbreviations - (Repr): method used for representation of process flow; (EA): whether it permits Ex-ante modelling; (EP): whether it permits Ex-post modelling; (Pu): whether Purpose can be specified; (Pr): whether Processing can be specified; (DS): if Data Sharing can be modelled; (Rp): if Recipients are associated with data sharing; (St): whether Data Storage occurs; (Rg): if provision of Rights can be modelled; and (LB): if Legal Basis can be associated with a process flow.

The table demonstrates that GDPRov supports all of analysed features and is the only one currently providing all of them. However, this analysis only takes into consideration abstract existence or provision of features and does not take into consideration context of an approach or its granularity. For example, while an approach may provide representation of data storage concepts, there are additional features such as storage duration, condition, form or medium, security, and policy which are also relevant in evaluation of GDPR compliance. These are highly dependant on individual use-cases and domains, and contain existing work which can be used to represent them such as Time ontology [184] for temporal annotations and ODRL ontology [185] for conditions and events as policies. Since the scope of GDPRov is limited to expression of information regarding activities in ex-ante and ex-post phases, representation of such granular attributes is relevant but not the primary focus within its scope and is therefore not considered in its evaluation or comparison with SotA.

Table 5.4: Comparison of GDPRov with SotA

Work	Repr	EA	EP	Pu	Pr	DS	Rp	St	Rg	LB
GDPRov	PROV-O,P-Plan	✓	✓	✓	✓	✓	✓	✓	✓	✓
SPECIAL	PROV-O	✓	✓	✓	✓	✓	✓	✓		
SPL+CitySPIN	PROV-O	✓	✓	✓	✓	✓	✓	✓		
MIREL	PWO	✓		✓	✓			✓	✓	
MRL+DAPRECO	PWO	✓		✓	✓			✓	✓	
BPR4GDPR		✓	✓	✓	✓	✓	✓			
Ujcich et al.	PROV-O		✓	✓	✓	✓	✓	✓	✓	✓
Lodge et al		✓		✓						
Tom et al	BPMN	✓			✓	✓	✓	✓	✓	
LUCE		✓	✓			✓	✓			
Sion et al		✓		✓	✓	✓	✓	✓		✓
privacyTracker		✓	✓			✓	✓			
Basin et al		✓		✓						
RestAssured				✓	✓	✓	✓	✓		

Regarding expression of information as being in either ex-ante or ex-post phase - GDPRov is the only approach to do so using existing ontologies of PROV-O and P-Plan based on a scientific workflow model which is useful in investigation of executions based on a plan and association of information between ex-ante and ex-post phases. The use of PWO [179] in MIREL project also follows a similar rationale though it does not provide the same extent of concepts and representations as GDPRov. In addition, utilisation of PROV-O as its base ontology enables information represented using GDPRov to be captured and recorded as provenance information using PROV-O or GDPRov itself. This provides capabilities for documenting evolution of a system as well as representing state of compliance at a given moment in time. This feature is shared by all approaches that utilise a provenance based ontology at their core, and especially the ones that utilise PROV-O as it is a well-recognised and adopted standard. The combination of PROV-O and P-Plan enables a flexible representation of activities by specifying their constituent steps and involved artefacts at an arbitrary level of granularity while providing annotations and classes to link these with GDPR. In this, GDPRov is unique and novel within SotA.

Based on this, GDPRov's novelty within SotA is based on it being one of few approaches using PROV-O to represent activities for GDPR compliance in both ex-ante and ex-post phases. GDPRov is also novel in its provision of concepts associated with GDPR and granularity afforded by use of PROV-O and P-Plan to link information in ex-ante and ex-post phases. Furthermore, GDPRov is one of few approaches to be available under an open and permissive license (CC-by-4.0) thereby enabling its use, adoption, and further evolution.

5.3.4.3 Application to external use-case from SPECIAL project

The SPECIAL project uses a scenario[45] to motivate their work and demonstrate use of developed technologies in their deliverables (D1.7 [186]) and peer-reviewed publications ([85]). In

[45]NOTE: the scenario explicitly mentions use of an immutable distributed ledger developed by the SPECIAL project to provide transparency and log accountability regarding metadata. This is omitted from the adapted use-case used to evaluate GDPRov.

this section, the scenario is adapted as an external use-case to evaluate GDPRov's suitability to express required concepts.

The use-case is summarised as follows with GDPR concepts added in parenthesis for relevance: Sue (Data Subject) buys a wearable appliance for fitness tracking from BeFit (Data Controller), and is presented with an informed consent request that describes collection of biomedical parameters such as heart rate (Personal Data) and how they will be processed, which are stored in BeFit's cloud and transmitted for purposes of: giving Sue feedback on her activity, such as calories consumption; and creating an activity profile that will be shared with other companies for targeted ads related to fitness - an optional purpose to which Sue opts-in. After two years, Sue starts receiving annoying SMS messages from a local gym that advertise its activities. Sue discovers following facts: (i) the gym has an activity profile referring to Sue, that, due to appliance's malfunctioning, reports that she is not doing any physical exercise; (ii) the gym received Sue's profile from BeFit, associated with a policy that allows the gym to send targeted ads to Sue based on the profile; (iii) BeFit built Sue's profile by mining data collected by appliance; and (iv) all these operations are permitted by consent agreement previously signed by Sue and BeFit. Using this information BeFit and the gym prove that they used Sue's data in accordance with Sue's given consent. Sue now asks both BeFit and the gym to delete all of her data.

The use-case is accompanied with information on its interpretation in terms of GDPR terminology [186] and its representation using SPECIAL vocabularies [85]. To represent the use-case using GDPRov, concepts used by SPECIAL are mapped or aligned to their closest relative concepts within GDPRov (see Table 5.5) with its RDF/Turtle representation provided in Listing 5. The SPARQL queries used to retrieve information depicted by statements (i) to (iv) in the scenario are provided in Listing 6. The RDF representation and SPARQL query utilised a simplified representation of the scenario to present only the essential fact for answering of questions.

Table 5.5: GDPRov concepts to represent external use-case from SPECIAL

Statement	**GDPRov concept**	**SPECIAL concept**
Sue	`DataSubject`	`DataSubject`
BeFit	`DataController`	`Controller`
Biomedical parameters, heart rate, calories consumption, activity profile	`PersonalData`	`Data`
Collect data	`DataCollectionActivity`	`Collect`
Provide feedback on activity	`Purpose`	`Purpose`
Give consent (opt-in)	`AcquireConsentActivity`	`ConsentAssertion`
Targeted ads related to fitness	`Purpose`	`Purpose`
Share data	`DataSharingActivity`	`Recipient`
Gym	`ThirdParty`	`Recipient`
Consent agreement	`GivenConsent`	`LogEntryContent`
Delete data	`DataDeletionActivity`	N/A
Withdraw consent	`WithdrawConsentActivity`	`ConsentRevocation`

Through the representation of concepts within the scenario and use of SPARQL, GDPRov

```
# Entities
:Sue a gdprov:DataSubject .
:BeFit a gdprov:DataController .
:Gym a gdprov:ThirdParty .
# Personal Data
:Biomedical_Parameters a gdprov:PersonalData .
:Activity_Profile a gdprov:PersonalData .

# Register with BeFit, given consent, and generate activity profile
:Registration a gdprov:Process .
:Sue_consent a gdprov:GivenConsent, gdprtext:LawfulBasisForProcessing .
:Collect_consent a gdprov:AcquireConsentActivity ;
    gdprov:isPartOfProcess :Registration .
    gdprov:collectedConsentFromAgent :Sue ;
    gdprov:generatedConsent :Sue_consent .
:Collect_data a gdprov:DataCollectionActivity ;
    gdprov:isPartOfProcess :Registration ;
    prov-o:wasInformedBy :Collect_consent ;
    gdprov:collectedDataFromAgent :Sue ;  # from Sue's device
    gdprov:generatedData :Activity_Profile .

# Share activity profile with Gym
:Targeted_ads_related_to_fitness a gdprov:Process ;
    gdprov:hasLegalBasis :Sue_consent .
:Share_data a gdprov:DataSharingActivity ;
    :sharedData :Activity_Profile ;
    :hasSharedDataWith :Gym .
# Gym receives Sue's activity profile from BeFit
:Collect_data_from_BeFit a :DataCollectionActivity ;
    gdprov:collectedDataFromAgent :BeFit ;
    gdprov:involvesAgent :Sue ;
    gdprov:refersToProcess :Targeted_ads_related_to_fitness ;
    gdprov:generatedData :Activity_Profile . # Gym copy of data

# Sue withdraws consent
:Withdraw_consent a gdprov:WithdrawConsentActivity ;
    prov:invalidated :Sue_consent .
# BeFit and Gym deleted the activity profile
:Delete_data a gdprov:DataDeletionActivity ;
    prov:invalidated :Activity_Profile ;
    prov:wasInformedBy :Withdraw_consent .
```

Listing 5: GDPRov representation of external use-case from SPECIAL

```
# Query (i)
# retrieves :Activity_Profile as personal data shared with Gym
# queried over Gym's records
SELECT ?personal_data
WHERE {
    ?personal_data a gdprov:PersonalData .
    ?activity gdprov:generatedData .
    ?activity (gdprov:collectedDataFromAgent|gdprov:involvesAgent) :Sue .
}

# Query (ii)
# retrieves :BeFit as data source
# retrieves :Targeted_ads_related_to_fitness as purpose
# queried over Gym's records
SELECT ?party, ?purpose
WHERE {
    ?activity gdprov:generatedData :Activity_Profile .
    ?activity gdprov:collectedDataFrom ?party .
    ?activity gdprov:refersToProcess ?purpose .
}

# Query (iii)
# retrieves :Sue as data source (as Sue's device)
# queried over BeFit's records
SELECT ?data_source
WHERE {
    ?activity_profile gdprov:generatedData :Activity_Profile .
    ?activity_profile gdprov:collectedDataFromAgent ?data_source .
}

# Query (iv)
# retrieves :Sue_consent as the legal basis for data collection and mining
# queried over BeFit's records
SELECT ?legal_basis
WHERE {
    {
        ?activity gdprov:generatedData :Activity_Profile .
    } UNION {
        ?activity gdprov:sharedData :Activity_Profile .
        ?activity gdprov:hasSharedDataWith :Gym .
    }
    ?activity (
            gdprov:hasLegalBasis|
            (gdprov:isPartOfProcess/gdprov:hasLegalBasis))
        ?legal_basis .
}
```

Listing 6: SPARQL queries using GDPRov for external use-case from SPECIAL

is shown to support representation and answering of questions within the scenario. While the scenario relies on use of SPECIAL's policy-based vocabulary and immutable distributed ledger to store and retrieve information regarding given consent and processing activities carried out by BeFit and the Gym, these have not been replicated as storage mechanisms are not relevant within the use-case as adapted for this work.

One important conclusion from the above exercise is regarding representation of given consent where SPECIAL vocabularies represent given consent as a policy using OWL2 [85]. In comparison, GDPRov does not represent information provided within consent requests and given consent, but instead records its activities and associates generated consent as an artefact representing a legal basis with purposes (specified using `gdprov:Process`) that rely on it. This is the consequence of GDPRov focusing on representation of activities which does not concern representation of consent. A consent-centric representation of this same scenario is presented in Section 5.4.5.3 which uses GConsent to represent the scenario as interactions between consent states.

Summary

GDPRov is part of the second major contribution of this thesis. It provides an ontological representation of ex-ante and ex-post activities associated with personal data and consent for GDPR compliance. It thus fulfils research objective $RO3(b)$ as outlined in Section 1.2. The use of GDPRov makes it possible to indicate plans associated with how personal data and consent is collected, used, stored, shared, and erased. It also enables representation of logs for activities that act over personal data and consent.

When GDPRov was first being developed (in 2016-2017), no other vocabulary was found that represented information about activities associated with GDPR compliance. The work presented as state of the art in Chapter 3 and demonstrating existence of approaches for representing information about GDPR processes were published after development and publication of GDPRov [66]. Of these approaches, some also utilise PROV-O to represent provenance of activities as found as presented in Section 3.7.3. The differentiating factor of GDPRov is in use of PROV-O and P-Plan as distinct ontologies representing ex-ante and ex-post phases of activities based on a scientific workflow model - which is novel within state of the art. Another differentiating factor is use of GDPRtEXT to define origin and relevance of concepts to their basis in GDPR.

Approaches within state of the art, such as SPECIAL (Section 3.2), demonstrate applicability of provenance vocabularies in maintaining, querying, and assessing provenance logs represented using PROV-O for GDPR compliance. While SPECIAL also provides ex-ante compliance assessment by using the same data model and logs it as a consent request instead of processing or execution [52], GDPRov expands further on use of provenance to include representation of plans or templates to indicate association between activities in ex-ante and ex-post phases of compliance.

5.4 GConsent - Ontology of Consent Information for GDPR Compliance

GConsent is a semantic web ontology for representing contextual information about consent based on requirements of GDPR compliance. GConsent aims to model context, state, and provenance of consent as an entity. Its scope is limited to consent as defined in GDPR and is intended towards assisting in modelling and management of information associated with compliance. It uses GDPRtEXT to denote origin and relevance of its concepts within GDPR.

GConsent is the outcome of applying the methodology presented in Section 5.1 to identify and represent information about consent and its life-cycle as required to determine compliance with GDPR. For this, information presented in Chapter 2 was used to identify validity of consent with requirements and compliance questions presented in Chapter 4 used as competency questions. The latest iteration of GConsent (v0.5) is published online[46] with its documentation under an open and permissive license of CC-by-4.0 and its code repository[47].

The design of GConsent was influenced by a real-world use-case for managing consent information based on GDPR compliance requirements, as mentioned earlier in Section 4.2.1. The design of ontology underwent several iterations based on whether it should model an association or dependency between purposes and processing operations associated with consent. The outcome, representing a decision and which is presented here, models separation between purpose and processing operations similar to other representations of consent within SotA.

5.4.1 Distinction with existing work in state of the art

Information about consent needs to be maintained and shared by multiple parties which includes data subjects who give consent, controllers who use it as legal basis, and authorities who evaluate its validity. GDPR requires information about consent across its life-cycle to be maintained, with an representation that is interoperable assisting all stakeholders in the compliance process - as outlined earlier in Section 4.1.

From existing work analysed in Chapter 3, the focus of approaches for consent is mostly on concept of 'given' consent i.e. consent provided by a data subject and used as legal basis by a controller. There is a lack of work regarding representing other 'states' of consent within its life-cycle as an entity or representation of agreement which are relevant to its use as legal basis in determination of processing of personal data and its compliance under GDPR. Examples of such states are 'not given', 'refused', 'withdrawn' which cannot be modelled in the same manner as 'given consent' as they do not reflect the same information as given consent, but are still relevant when associated with a particular instance of processing. The state of a consent reflects its status for use as legal basis and is also relevant in management of consent information from an organisation's perspective.

Apart from the notion of states, existing approaches also lack modelling representations for events such as delegation, and associations with third parties regarding consent which have an effect on its validity regarding compliance. GConsent aims to fill this gap, and

[46] https://w3id.org/GConsent
[47] https://github.com/coolharsh55/GConsent/

therefore to provide novelty and contribution by representing a more cohesive and complete representation of information associated with consent for GDPR.

5.4.2 Relationship with GDPRov

GConsent builds upon and is complimentary to the representation of consent in GDPRov. The definition of consent as an entity involved in activities is sufficient to express its life-cycle and provenance by using GDPRov. This includes ex-ante representation of information provided to collect consent and its subsequent agreement by an individual to produce given consent - which is then used within activities as a legal basis and may be modified, withdrawn, or revoked - signalling its effective end of life-cycle. While GDPRov is sufficient to represent these states of consent as an entity along with information about activities acting on it, the primary focus of GDPRov is for expressing information regarding its association with activities. Managing consent as a legal basis involves consideration of information such as purpose of processing, recipients, and contextual information such as medium of provision and collection, and situations such as delegation - which are not modelled in GDPRov.

GConsent aims to provide a consent-centric representation of these information categories by providing concepts relevant to resolution of valid consent as defined by requirements of GDPR compliance. In this, use of provenance concepts show an overlap with GDPRov. This is resolved through differing scopes of the two ontologies, where concepts defined within GDPRov can also be defined using corresponding concepts in GConsent and vice-versa. An example of such cohesive usage is demonstrated through application of both GDPRov and GConsent for querying and validation of information in Chapter 6. By having separation between GDPRov and GConsent, the ontologies are modular in their scope and concepts, and provide an adopter with choice regarding inclusion of semantics represented by each ontology.

5.4.3 Requirements Gathering and Establishment of Competency Questions

The scope of consent as represented within GConsent is limited to definition of consent as provided by Article 4-11[48] of GDPR. Other special cases of consent not included within GConsent consist of consent defined by Article 9 regarding use of special categories of personal data, Recital 33 regarding use of personal data for scientific research, and Article 8 along with Recital 38 regarding use of children's personal data. These were not included within the scope due to their additional requirements and complexity regarding interpretation and representation using semantic web. Additionally, current lack of real-world use-cases reflecting how these types of consent would function and legal guidance on their compliance requirements is noticeably absent in context of GDPR.

[48] In this definition, consent is expressed as the indication of a data subject's wishes regarding processing of their personal data. However, this is a legal definition rather than a semantic one as it essentially defines the set of conditions required to be satisfied by some consent before it is considered valid under GDPR. Additionally, 'consent' as a social concept has a pre-defined meaning based on its use in the social context. Therefore, referring to consent in the context of GDPR does not only mean given consent but includes all information and states associated with consent which can then be evaluated to assess whether it fits the definition of consent as a legal basis. Technical approaches can use 'consent' to indicate the given consent or the set of all consent states within its life-cycle.

GConsent is primarily based on notion of consent as a legal basis under Article 6 of the GDPR. The conditions defined in Article 7, Recital 42, and Recital 43 provide requirements for consent to be considered or determined valid as per requirements of GDPR. The Data Controller bears burden of demonstrating proof and satisfaction of requirements for consent to be considered valid as per Recital 42. This requires demonstrable proof that data subject provided consent and that it was valid as per obligations specified in GDPR.

For consent to be informed it is necessary to provide information to data subjects which includes specific purposes of processing the personal data. GDPR also provides data subjects with right to modify or withdraw consent as defined in Article 7-3. When consent is withdrawn, processing carried out done prior to withdrawal is considered valid under given consent in effect as legal basis during that period of time.

Through this, a rudimentary summary of information or attributes associated with 'consent' as defined by GDPR is expressed as:

- Data Subject the consent is about
- Personal Data associated with consent
- Processing operations or categories the consent is about, with data storage and data sharing having additional information requirements
- Purposes the consent is about
- Entity/Agent/Actor the consent is provided to
- Recipients of data or categories of recipients if any

These attributes are sufficient to provide a simplified representation of consent, and are used in existing approaches within state of the art - such as model of consent in SPECIAL vocabularies [187]. However, these attributes are not sufficient by themselves to determine validity of consent as they lack information about context the consent was given in as well as changes to its state.

Therefore, further attributes associated with consent are identified and expressed as:

- Entity/Agent/Actor that provided consent - relevant in case of delegation
- Entity/Agent/Actor that revoked, withdrew, or invalidated consent - relevant in case of delegation, and authoritative actions such as by regulators or courts
- Status of consent at a given period in time
- Contextual information regarding request and giving of consent such as - location, medium, timestamp, expiry
- Contextual information regarding revocation, withdrawal, and invalidation of consent such as - location, medium, timestamp, expiry

In addition to these, provenance information regarding how consent was requested and obtained is also important. Specifically, information about specific processes and artefacts used in provision of request for consent should be recorded as they must satisfy GDPR qualitative requirements - such as request being clearly stated and being unambiguous.

To derive required concepts, competency questions were identified from compliance questions presented in Section 4.2 pertaining to given consent (`CMQ35-CMQ69`) and change in consent state (`CMQ70-CMQ87`). These refer to information regarding consent (e.g. `CMQ35-CMQ40`), how consent was created/given/changed/invalidated (e.g. `CMQ41-CMQ52`), context of how consent was created/given/invalidated (e.g. `CMQ53-CMQ56`), and third parties associated with consent (e.g. `CMQ57-CMQ58`).

5.4.4 Ontology Description & Application

5.4.4.1 Core Concepts

Core concepts and relationships in GConsent describe common and primary attributes associated with consent. In this case, 'consent' by itself does not refer only to state of 'given consent' but also stands as a representation of 'consent' as an entity whose state is unknown or is refused, withdrawn, or invalidated by Data Subject, Controller, or an authority such as courts. This definition of consent is based on managing consent as a data entity rather than as a semantic concept pertaining to an agreement by an individual. Core concepts are associated with consent in all its states and refer to information necessary to express what the consent is about. This comprises of the 5 attributes visualised in Figure 5.10 - Data Subject, Personal Data, Purpose, Processing, and Status.

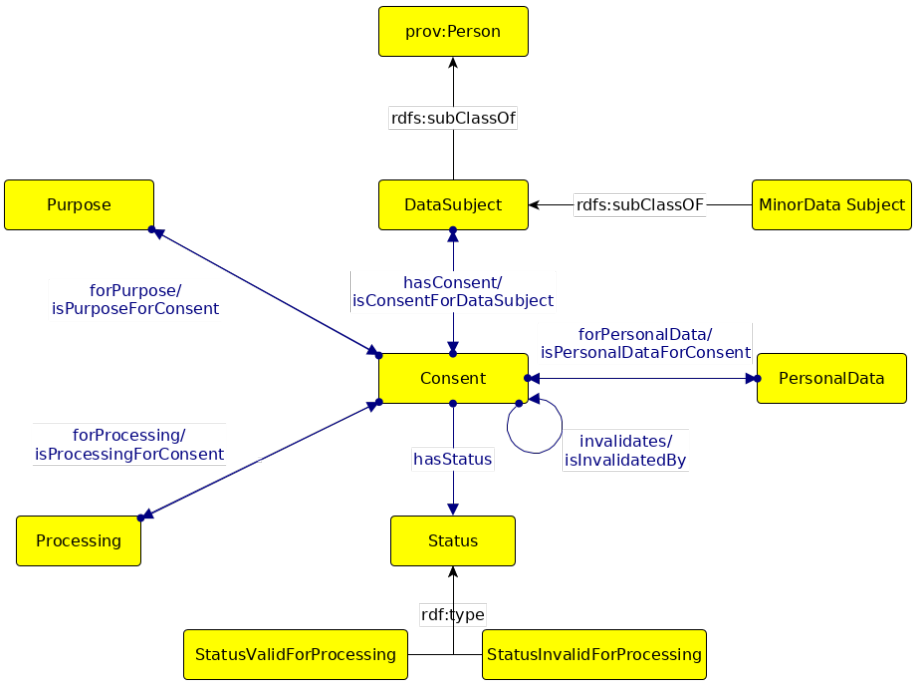


Figure 5.10: Core concepts in GConsent [71]

`DataSubject` is natural individual consent is associated with as an agreement of their choices. This individual may or may not be the same entity that gave consent - as in case of parent or guardian giving consent for a child or as an act of delegation. `DataSubject` class is defined as a subclass of `prov:Person`, and with subclass `MinorDataSubject` to denote a data subject that is legally a minor or a child. Each instance of `Consent` must be associated with one and only one `DataSubject`, and any further changes or modifications to a state of consent will continue to be associated with the same `DataSubject`. `PersonalData` is a set of personal data associated with consent. Where multiple personal data are associated with a single instance of consent, it is interpreted to mean union of these sets of personal data. Similarly, multiple `Purpose` and `Processing` associated with a consent are also to be interpreted as union rather than intersection. The 'status' or 'state' of consent indicates suitability of using that specific instance of consent as a legal basis for processing of personal data as defined by associated attributes.

`Purpose` and `Processing` are concepts that have semantic meaning based on their use within GDPR. 'Processing' is defined by Article 4-2, while 'Purpose' has no specific definition provided but can be summarised as intent or aim of why the set of personal data is needed or to be used for. In practice, purpose is generally defined at a higher abstract level, and often encompasses several types or categories of data. An example of this is a privacy policy specifying 'account information' and 'location of service use' - which are data categories, that are 'collected' and 'used' - which are processing operations on personal data, 'to ensure security of the account' - which is the purpose personal data will be processed for. The relation between a purpose and its associated processing operations is quite opaque when considered for purposes involving one or more processing operations. Based only on the description, it is difficult to determine which processing operations a purpose entails and vice versa, and their usage may not always be implied or commonly understood. Therefore, GConsent provides purpose and processing as self-declarative high-level concepts which can be extended with additional information for granularity and transparency.

5.4.4.2 Context of Consent

The context of consent refers to attributes such as location or time when instance of consent was created, invalidated, generated, changed, modified, given, or recorded. GConsent provides concepts for expressing location, medium, and timestamp to indicate instant of creation or invalidation along with capturing 'expiry' of consent as either an instant of time or a duration using Time vocabulary [184]. The context also represents how consent was 'provided' by a Person or Data Subject or Delegation. The provided contexts in GConsent are visualised in Figure 5.11.

Context is associated with an instance of consent using generic property `hasContext`, with specialised properties extending it to indicate provision, expiry, location, time, and medium. Additional contexts can be represented and associated by extending `hasContext` in a similar manner.

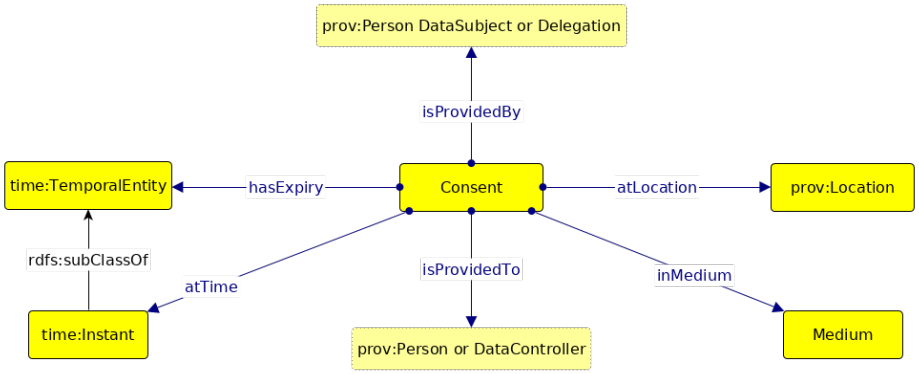


Figure 5.11: Concepts for representing context of consent in GConsent [71]

5.4.4.3 Consent States

The state of consent determines suitability of its usage as a legal basis in processing of personal data. From a compliance perspective, there are only two categories of states - one which permits legal processing of personal data, and other being insufficient or prohibitive

for processing. GConsent represents these concepts by sub-classing `Status` as `Status-ValidForProcessing` and `StatusInvalidForProcessing` to indicate use of a consent instance as valid or invalid legal basis as depicted in Figure 5.12. Instances provided to represent states of valid consent to indicate legal processing include - explicitly given, implicitly given, and given by delegation. Instances provided that represent invalid states of consent to indicate processing should not be carried out include - unknown, not given, withdrawn, expired, invalidated, refused, and requested.

The use of state refers to tracking consent of a data subject from a legal perspective, and is aimed to aid in management of consent as an entity. For example, 'unknown' reflects a situation where status of consent is not known - which can occur when importing consent information from another source. This is distinct from 'not given' which indicates an offer has been made for obtaining consent but a data subject has not yet provided any actionable response that could indicate acceptance or refusal - which are themselves represented by states 'given' and 'refused' respectively. For meeting obligations and requirements of GDPR compliance, it is not necessary to represent consent instances with states such as unknown or refused. GConsent provides them for practical management of consent information where a controller may wish to track consent status of its processing operations throughout its life-cycle.

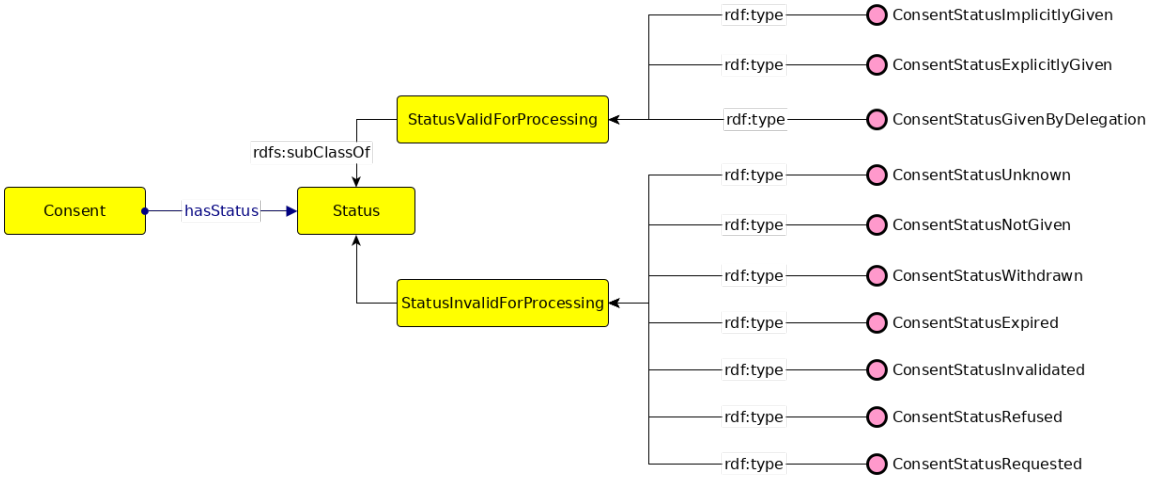


Figure 5.12: Concepts representing state/status of consent in GConsent [71]

GDPR requires keeping track of state change for consent - for example when status changes from given to withdrawn or when consent is invalidated because by a controller or legal authority. Whenever a consent status changes, this results in a new consent instance being created, which also assists in capturing context of the consent (such as time instant). This leads to a chain of consent instances, where 'latest' consent is at 'end' of this chain and indicates most recent operation regarding consent states. It is vital to record such provenance to demonstrate past processing was in compliance with state of consent at that point in time and to show changes to consent as part of its life-cycle.

5.4.4.4 Example Use-Case

The documentation of GConsent provides example applications in four use-cases to demonstrate how information can be represented, which are - (i) change in consent state, (ii) capturing given consent, (iii) capturing consent given via delegation, and (iv) capturing consent when data is shared with a third party. The fourth use-case is presented here to demonstrate application of GConsent and use of its concepts to represent information towards GDPR compliance.

The example, visually represented in Figure 5.13, shows association of a third party in role of a data processor[49] with whom data is shared for purposes of advertising. The association is captured by instance `ex:AdvertisingArrangement` of type `prov:Association`, and has `ex:AdPartner` defined as a `gdprov:Processor` defined with role as `gdprtext:Processor`. It is also possible to list out specific arrangement for this association using `prov:hadPlan` property and a `gdprov:Process` instance to list specific steps and entities involved in data sharing arrangement.

The example serves to demonstrate practical use of GConsent in representing information about consent, where PROV-O is used to specify relationships with a Processor. GConsent can be combined or supplemented with other ontologies to define such associations and practical reflections of data sharing agreements between parties. The defined instance of consent in example enables a Controller to track state of consent as the data subject is provided with choice of whether to agree to this arrangement or to refuse it, where upon agreement the option to exercise right to modify and withdraw consent is also provided.

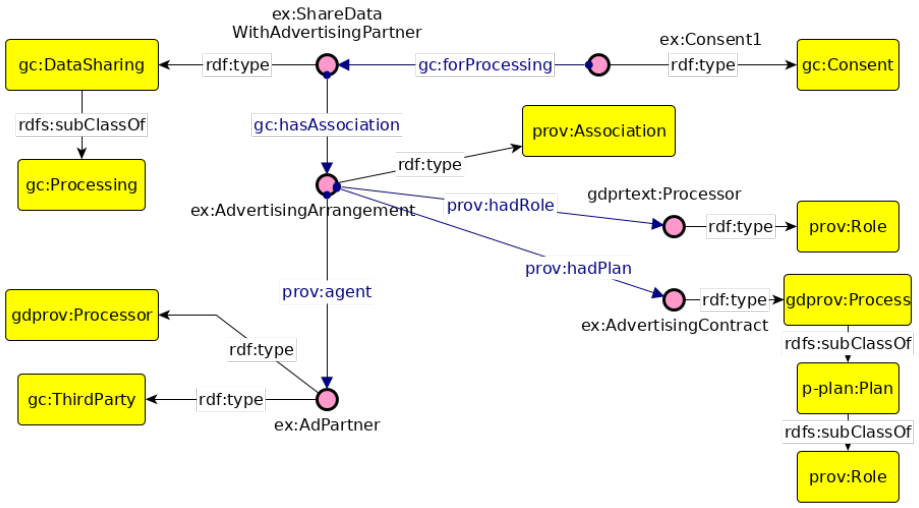


Figure 5.13: GConsent representation of use-case involving third party data sharing [71]

5.4.5 Evaluation

GConsent as an ontology was evaluated regarding its capability to express information about consent using the methodology outlined in Section 5.1. This was an iterative pro-

[49] Under GDPR, a processor is not considered a third party, but has its own defined role as an entity associated with the Controller. However, from a lay person's perspective, the individual is the first party, the Controller is the second party, and any other entity is a third party. GConsent reflects this use in its structuring of entities where a Processor is considered a special type (sub-class) of Third Party.

cess where the ontology was tested and modified to accommodate the requirements of the competency questions. Changes were made to the ontology where information was found to be missing or incorrectly modelled. In particular, the iterations consisted of the degree and design of representing a dependency between purposes and processing operations associated with consent. These were ultimately rejected with the final iteration modelling purpose and processing independent of each other to provide greater granularity and reuse of these concepts. An application of GConsent towards representation of consent information for a real-world website in the application of SHACL to validate information for GDPR compliance is presented in Section 6.2.

GConsent was published as a peer-reviewed publication [71] in Extended Semantic Web Conference (ESWC) Ontologies and Reasoning Track. As ESWC is a top-tier semantic web conference with a rigorous review process, acceptance of GConsent demonstrates its contribution as a semantic web resource. Along with this, the documentation of GConsent, available online, also provides extensive information about the ontology and its potential applications. It also provides a brief comparison of the ontology with relevant approaches within state of the art. GConsent, along with Consent Receipt standard [170], had a direct impact on design and development of consent information within DPV. In particular, GConsent provided concepts for expression of consent based on the GDPR and competency questions for integrating those with rest of DPV. As of February 2020, GConsent does not currently have any citations (excluding self-citations) given recency of its publication.

5.4.5.1 Fulfilment of Competency Questions

Assessment of the extent to which GConsent provides concepts and relationships to answer competency questions is presented in Table 5.6. In these, the PROV-O and Time ontologies are used in conjunction with GConsent to represent provenance and temporal information about consent and changes to it. PROV-O is also used to capture association and roles of entities in activities associated with consent. Based on these, GConsent satisfies requirements of providing information for answering compliance questions regarding consent and thus fulfils research objective $RO3(c)$.

Table 5.6: Concepts in GConsent for answering competency questions

CQ	Question	Concepts	Properties
Questions about consent			
CMQ35	Who is the consent about?	*DataSubject*	*isConsentForDataSubject*
CMQ36	What type of Personal Data are associated with the Consent?	*PersonalData*	*forPersonalData*
CMQ37	What type of Purposes are associated with the Consent?	*Purpose*	*forPurpose*
CMQ38	What type of Processing are associated with the Consent?	*Processing*	*forProcessing*
CMQ39	What is the Status of Consent?	*Status*	*hasStatus*
CMQ87	Is the current status valid for processing?	*StatusValidForProcessing*, *StatusInvalidForProcessing*	*hasStatus*
CMQ46	Who is the consent given to?	*prov:Person*, *DataController*	*isProvidedTo*

(Cont'd on following page)

Concepts in GConsent for answering competency questions (cont'd)

CQ	Question	Concepts	Properties
Questions about how the consent was created/given/acquired/changed/invalidated			
CMQ42, CMQ76	Who created/gave/acquired/invalidated the consent?	*DataSubject, Delegation*	*isProvidedBy*
CMQ41, CMQ77	If consent was created/gave/acquired/invalidated through Delegation, who acted as the Delegate?	*prov:Person, Delegation*	*prov:agent*
CMQ43	If consent was created/gave/acquired/invalidated through Delegation, what was the role played by Delegate?	*prov:Role*	*prov:hadRole*
CMQ44	If consent was created/gave/acquired/invalidated through Delegation, how was the delegation executed?	*prov:Activity*	*prov:hadActivity*
Questions about the context of how consent was created/gave/acquired/invalidated			
CMQ53, CMQ84	What is the location of associated with consent?	*prov:Location*	*atLocation*
CMQ54, CMQ85	What is the medium associated with consent?	*Medium*	*inMedium*
CMQ55, CMQ86	What is the timestamp associated with the consent?	*time:Instant*	*atTime*
CMQ56, CMQ87	What is the expiry of the consent?	*time:TemporalEntity*	*hasExpiry*
CMQ82	What artefacts were shown when consent was acquired/changed/created/invalidated?	*prov:Entity*	*prov:used*
Questions related to Third Party associated with the consent			
CMQ57	Is the purpose or processing associated with a third party?	*prov:Association, ThirdParty*	*hasAssociation, prov:agent*
CMQ58	What is the role played by the third party in the purpose or processing?	Role	prov:hadRole

5.4.5.2 Comparison with SotA

Existing approaches regarding consent were presented and analysed in Section 3.7.4, with an observation about lack of approaches modelling consent as required for GDPR compliance. Table 5.7 demonstrates a comparison of GConsent with SotA based on attributes used in this analysis. Column headings indicate representation of information within an approach and are abbreviated to indicate - Personal Data (PD), Purpose (Pu), Processing (Pr), Data Sharing (Sh), Data Storage (St), Recipients (Rp), Data Source (S), Withdrawal of consent (W), Delegation (D), Visualisation (V), Significant effects of processing (SE), (Ct): Context (Ct), Types or States (T). A check mark (✓) indicates the approach provides or models that information category, and a blank cell indicates that the approach does not provide representation for that information or that there is no open and public information available regarding its provision.

The table demonstrates contributions of GConsent to state of the art. Compared to SotA,

GConsent provides novel contributions for representation of consent for GDPR compliance and thus extends state of the art. In particular, depiction of delegation is more detailed and provides representation of information based on compliance requirements of GDPR. As the table depicts, GConsent is currently the only approach that models delegation based on its potential relevancy to evaluation of GDPR compliance. GConsent is also novel in provision of consent states which enable documenting of information from a controller's perspective regarding evolution of consent throughout its life-cycle. This is useful for management of consent as an entity in an information management system such as a database. The SotA usually limits consent state to given or withdrawn without consideration to its other states within its life-cycle as an entity.

Table 5.7: Comparison of GConsent with SotA

Work	PD	Pu	Pr	Sh	St	Rp	S	W	D	SE	Ct	T
GConsent	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
SPECIAL	✓	✓	✓	✓	✓	✓		✓				
SPL+CitySPIN	✓	✓	✓	✓	✓	✓		✓				
Lodge et al	✓	✓										
Peras	✓	✓	✓	✓	✓			✓				
Coletti et al	✓	✓					✓	✓				
AdvoCATE	✓	✓			✓	✓				✓	✓	
RestAssured	✓	✓	✓	✓	✓	✓						
OPERANDO	✓	✓	✓	✓		✓						
PoSEID-on	✓					✓						
MHMD	✓											
DECODE	✓	✓			✓							
Consent Receipt	✓	✓									✓	✓

5.4.5.3 Application to external use-case from SPECIAL project

The use-case described earlier in Section 5.3.4.3 is used here to demonstrate use and sufficiency of GConsent in management of consent information. The use-case concerns the scenario where a data subject named Sue gives consent to BeFit company for sharing her activity profile with other companies for receiving targeted ads. She later receives ads from a local Gym and investigates to find that the Gym is using her activity profile shared by BeFit - and that this activity is consistent with her previously given consent. She then withdraws her consent and asks both companies to delete her data.

The representation of this use-case using GConsent consists of using its concepts to represent information, and then utilise SPARQL to answer queries - similar to the exercise in Section 5.3.4.3 for GDPRov. Table 5.8 presents concepts for representing use-case using GConsent, GDPRov, and SPECIAL vocabularies. The corresponding RDF representation using GConsent is provided in Listing 7 with queries for deriving answers to questions (i) - (iv) in use-case provided in Listing 8.

Table 5.8: GConsent concepts to represent external use-case from SPECIAL

Statement	GConsent	GDPRov	SPECIAL
Sue	DataSubject	DataSubject	DataSubject
BeFit	DataController	DataController	Controller
Biomedical parameters, heart rate, calories consumption, activity profile	PersonalData	PersonalData	Data
Collect data	Collection Of PersonalData	Data Collection Activity	Collect
Provide feedback on activity	Purpose	Purpose	Purpose
Give consent (opt-in)	N/A	Acquire Consent Activity	ConsentAssertion
Targeted ads related to fitness	Purpose	Purpose	Purpose
Share data	Sharing Of Personal Data	Data Sharing Activity	Recipient
Gym	ThirdParty	ThirdParty	Recipient
Consent agreement	Consent Status Explicitly Given	GivenConsent	LogEntryContent
Delete data	Deletion Of Personal Data	Data Deletion Activity	N/A
Withdraw consent	N/A	Withdraw Consent Activity	ConsentRevocation
Withdrawn consent	Consent Status Withdrawn	N/A	N/A

From the above representation and queries, GConsent demonstrates use of terms (purpose, sharing, etc.) closer to those of SPECIAL as compared to GDPRov's representation in Section 5.3.4.3. At the same time, an inability to represent information about activities such as how data was collected or shared is also evident since GConsent does not represent them while GDPRov does. From a consent perspective, consent 'records' represented using GConsent are more clear and concise in terms of what consent is related to, and how it was withdrawn. The above representation therefore demonstrates GConsent's use in managing consent information, with SPARQL queries used to retrieve answers to questions pertaining to use of Sue's personal data within the use-case.

Summary

GConsent is an ontology for representation of consent and its associated information for GDPR compliance. It fulfils research objective $RO3(c)$ and along with GDPRov forms the second major contribution of this thesis. GConsent has been published in a peer-reviewed publication and is available online as an open and reusable resource along with an extensive and descriptive documentation.

GConsent is currently the only approach within state of the art to provide representations of attributes of consent and its states based on requirements of the GDPR. GConsent thus represents a novel representation of consent based on GDPR and provides concept of states for practical management of consent information from a controller's perspective. It also provide detailed information representation regarding context of consent which enables documenting information required for evaluating the validity of consent under GDPR compliance requirements.

```
# Entities
:Sue a gc:DataSubject .
:BeFit a gc:DataController .
:Gym a gc:ThirdParty .
# Personal Data
:Activity_Profile a gc:PersonalData .
# Purpose
:Targeted_ads_related_to_fitness a gc:Purpose .

# Sue gives consent to BeFit
:Consent1_registration a gc:Consent ;
        gc:isConsentForDataSubject :Sue ;
        gc:isProvidedToController :BeFit ;
        gc:forPurpose :Targeted_ads_related_to_fitness ;
        gc:forProcessing gc:CollectionOfPersonalData,
                gc:ShareDataForTargetedAds ;
        gc:forPersonalData :Activity_Profile ;
        gc:hasStatus gc:ConsentStatusExplicitlyGiven .

# BeFit shares data with Gym
# assumed similar 'policy' structure as SPECIAL
:ShareDataForTargetedAds a gc:DataSharing ;
        gc:involvesThirdParty :Gym .
        gc:sharesDataWithThirdParty :Gym .

:Consent_info_shared_by_BeFit_with_Gym a gc:Consent ;
        gc:isConsentForDataSubject :Sue ;
        gc:isProvidedTo :BeFit ;
        gc:forPurpose :Targeted_ads_related_to_fitness ;
        gc:forProcessing gc:UseOfPersonalData ;
        gc:forPersonalData :Activity_Profile ;
        gc:hasStatus gc:ConsentStatusExplicitlyGiven .

# Sue withdraws consent
:Consent2_withdraw a gc:Consent ;
        gc:isUpdatedConsentFor :Consent1_registration ;
        gc:forPurpose :Targeted_ads_related_to_fitness ;
        gc:forProcessing gc:CollectionOfPersonalData,
                gc:ShareDataForTargetedAds ;
        gc:forPersonalData :Activity_Profile ;
        gc:hasStatus gc:ConsentStatusWithdrawn .
```

Listing 7: GConsent representation of external use-case from SPECIAL

```
# Query (i)
# retrieves :Activity_Profile as data shared for Sue
# queried over Gym's records
SELECT ?personal_data
WHERE {
        ?_ gc:isConsentForDataSubject :Sue .
        ?_ gc:forPersonalData ?personalData .
}

# Query (ii)
# retrieves :BeFit as data source
# retrieves :Targeted_ads_related_to_fitness as purpose
# queried over Gym's records
SELECT ?party, ?purpose {
        ?_ gc:isConsentForDataSubject :Sue .
        ?_ gc:forPersonalData :Activity_Profile .
        ?_ gc:isProvidedTo ?party .
        ?_ gc:forPurpose ?purpose .
}

# Query (iii)
# retrieves :CollectionOfPersonalData as the processing operation
# from which it needs to be inferred that data is collected from Sue
# queried over BeFit's records
SELECT ?data_processing {
    ?_ gc:forPersonalData :Activity_Profile .
    ?_ gc:forProcessing ?data_processing .
}

# Query (iv)
# retrieves :Consent1_registration as satisfying
# conditions for data sharing with Gym
# queried over BeFit's records
SELECT ?consent
WHERE {
        ?consent a gc:Consent .
        ?consent gc:isConsentForDataSubject :Sue .
        ?consent gc:forPersonalData :Activity_Profile .
        ?consent (gc:forProcessing/gc:sharesDataWithThirdParty) :Gym .
}
```

Listing 8: SPARQL queries using GConsent for external use-case from SPECIAL

5.5 Data Privacy Vocabulary (DPV)

The Data Protection Vocabulary [78] is a semantic web ontology for representing information about personal data handling based on legal requirements such as those for GDPR compliance. It is the outcome of work done by W3C Data Privacy Vocabularies and Controls Community Group (DPVCG) which consists of collaboration between a community of academics, researchers, industry stakeholders, and legal experts as initially described in Section 1.4.6. DPVCG aims to work towards establishment of interoperable standards regarding representing information about personal data processing for which there are currently no existing standards.

The DPVCG was initiated as part of SPECIAL project [70], and therefore bears close association and alignment with SPECIAL vocabularies. In particular, SPECIAL core vocabulary was used as the basis to create the DPV core vocabulary, which provides compatibility between DPV and SPECIAL vocabularies and frameworks.

DPV reflects a community consensus in its representation of information regarding data protection and personal data processing. While being a generic vocabulary, much of its design is based on and reflected by requirements of GDPR. DPV, and by extension DPVCG, reflect an ongoing effort to provide practical and useful semantic representations of information in an open, interoperable, and machine-readable form.

5.5.1 Relevance of DPV to this thesis

The ontologies presented in this thesis as research contributions - GDPRtEXT, GDPRov, and GConsent - were part of state of the art analysed by DPVCG in its methodology [78]. In addition, by being an active member of DPVCG and a contributor in creation of DPV, the author of this thesis has applied the experience of developing research presented in this thesis and influenced design and modelling of information within DPV.

A peer-reviewed publication of DPV [78] presents its creation methodology and concepts where the author of this thesis was a co-first author. In addition to these, the vocabulary specification published online lists the author of this thesis as an co-editor and author of the ontology. The deliverable D6.5 [70] of SPECIAL project presents work of DPVCG and describes DPV based on its peer-reviewed publication [78] - where the author of this thesis was also a co-lead author for the deliverable.

Owing to involvement of the author and overlap between DPV and the research question and developed ontologies in this thesis, the DPV is presented here as an external research contribution influenced by research presented in this thesis. This section describes DPV and compares it with ontologies presented in this thesis to provide an extent of their similarity and overlap. It demonstrates differences in representation of information, scope, and methodology and their complimentary nature in representing information for GDPR compliance.

5.5.2 Overview of DPV

5.5.2.1 Description of Data Privacy Vocabulary

The DPV ontology is published in the W3C namespace http://w3.org/ns/dpv with its documentation and uses the namespace prefix `dpv`. Its current iteration (v0.1 28 November 2019) provides classes and properties to annotate and categorise information about legally compliant personal data handling. In this context, personal data handling refers to all operations associated with processing of personal data and its management - including organisational measures which indirectly affect processing.

The DPV is a pseudo-modular ontology with a set of core concepts referred to as '*Base Ontology*' and modular extensions further expanding each concept within the base ontology as a taxonomy. The base ontology represents top-level classes for defining a policy of legal personal data handling. The core concepts defining the base ontology are visualised in Figure 5.14 and consist of personal data category, processing, purpose, legal basis, data controller, recipient, data subject, technical and organisational measures, with the top-level concept of personal data handling which ties them together.

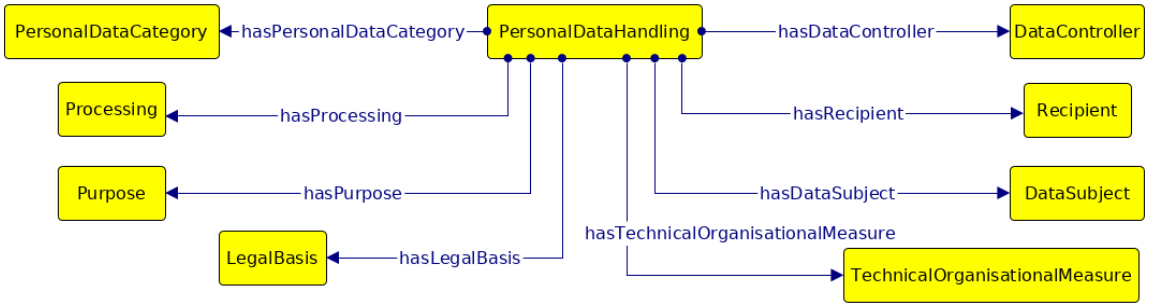


Figure 5.14: Core concepts in DPV [78]

5.5.2.2 Personal Data Categories

DPV uses the taxonomy provided by EnterPrivacy[50] to define a broad hierarchy of personal data categories based on nature of information (financial, social, tracking) and to its inherent source (internal, external). In addition to these, the class `dpv:SpecialCategoryOfPersonalData` represents categories that are 'special' or 'sensitive' based on GDPR's Article 9.

These personal data categories can be further extended using the sub-class mechanism to depict specialised concepts such as 'likes regarding movies'. Sub-classing also enables representation of specific contexts such as derivation of personal data as represented by class `dpv:DerivedPersonalData`. This is useful to represent practical representation of personal data categories such as inference of opinions from social media. Similar classes can be additionally added to specify contexts such as use of machine learning, accuracy, and source. The aim of providing such high-level concepts is to provide a sufficient coverage of abstract categories of personal data which can be extended using subclass mechanism to represent concepts used in real-world.

[50] https://enterprivacy.com/wp-content/uploads/2018/09/Categories-of-Personal-Information.pdf

5.5.2.3 Purposes

Purposes in DPV are organised hierarchically using the sub-class mechanism to represent high-level and generic purposes of data handling. Purposes provided in DPV include service provision, R&D, commercial interest, security, service optimisation, and service personalisation. These are further extended to provide a total of 31 generic purposes. These may be extended by further sub-classing to create more specific purposes as applicable to a scenario. As GDPR requires a specific purpose to be declared in an understandable manner, additional purposes can be created as sub-classes of one or several `dpv:Purpose` categories to make them as specific in a use-case as possible. Purposes can be restricted to specific `contexts` using the class `dpv:Context` and property `dpv:hasContext`. Purposes can also be restricted to a specific `business sector` using the class `dpv:Sector` and property `dpv:hasSector`.

5.5.2.4 Processing Categories

DPV provides a hierarchy of processing categories based on requirements of regulations such as GDPR. DPV defines top-level classes to represent following broad categories of processing - Disclose, Copy, Obtain, Remove, Store, Transfer, Transform, and Use. Each of these are further expanded using sub-classes to provide 33 processing categories, which includes terms defined in definition of processing in GDPR (Article 4-2). DPV provides properties with a boolean range to indicate nature of processing regarding `Systematic Monitoring`, `Evaluation or Scoring`, `Automated Decision-Making`, `Matching or Combining`, `Large Scale processing`, and `Innovative use of new solutions` - as these affect assessment of legal data processing under GDPR.

5.5.2.5 Technical and Organisational Measures

GDPR Article 32 requires implementing appropriate measures by taking into account state of the art, costs of implementation and - nature, scope, context, and purposes of processing - as well as risks, rights and freedoms. These are represented as technical and organisational measures in DPV. Examples include pseudo-anonymisation and encryption of personal data, ability to restore availability and access to personal data in a timely manner in event of a physical or technical incident, and a process for regularly testing, assessing and evaluating effectiveness of technical and organisational measures for ensuring security of processing. The generic property `dpv:measureImplementedBy` enables referencing implementation measures as comments or IRIs. The class `StorageRestriction` provides expression of measures used for storage of data with two specific properties provided for storage location and duration restrictions.

5.5.2.6 Consent and Legal Bases

DPV provides `dpv:LegalBasis` as a top-level concept to represent various legal bases that can be used for justifying processing of personal data. The definition of a 'legal basis' is based on justification for processing which has a provision in law. The concept itself is not based on any specific jurisdiction, but needs to be interpreted in terms of legal bases defined and provided by laws applicable within a jurisdiction.

For GDPR, which is a EU specific law and therefore is not binding in interpretation of legal bases across other jurisdiction, DPV provides legal bases specific to GDPR as a separate aligned vocabulary under https://www.w3.org/ns/dpv-gdpr and namespace (prefix: `dpv-gdpr`). This vocabulary defines legal bases provided by Articles 6 and 9 of GDPR to represent legal justification for processing personal data.

Consent as a special case of legal basis provided by GDPR is provided with additional properties and classes within the core DPV vocabulary to reflect information requirements associated with its validity as a legal basis. The concepts associated with consent provide terms to describe consent provision, withdrawal, and expiry. The structure of these was adapted from an analysis of existing work regarding Consent Receipt [170] and GConsent [71] with the intention to enable documenting attributes associated with consent which can demonstrate and evaluate its validity based on requirements of GDPR.

5.5.3 Comparing DPV with GDPRtEXT, GDPRov, and GConsent

DPV has several commonalities with ontologies presented in this thesis arising from a common aim of representing information for legal compliance with laws such as GDPR. However, the aim and granularity of representing information about all attributes relevant to processing of personal data differentiates DPV from the scope of ontologies presented in this thesis. Given the high-level and abstract nature of DPV in its concepts and more granular and representative focus of developed ontologies, there is a possibility of aligning or combining the two based on identifying commonality of concepts. While there have been no efforts to carry out an exercise to combine or align the developed ontologies with DPV, their comparison as presented here demonstrates the possibility and applicability of such an approach.

5.5.3.1 Representing information about GDPR concepts

GDPRtEXT provides a linked data version of GDPR and a SKOS glossary of concepts associated with GDPR compliance, which can be used to link information to clauses and concepts of GDPR. This has been used by GDPRov and GConsent to define the source of its concepts within GDPR.

DPV, whose concepts represent generic legal terms, does not link its concepts to GDPR except in cases where a defined concept was directly taken from a definition provided by GDPR. In such cases, it uses the URI format prescribed by EU Publications Office to indicate specific clause of GDPR. The URI format[51] is similar in its structuring of contents with GDPRtEXT and is based on an upcoming iteration[52] of ELI vocabulary which will be used in all EU published legislations to offer granular linking to their clauses. Since the format prescribed by EU Publications Office is authoritative in its nature, the links provided by GDPRtEXT need to be aligned or replaced with those defined using the newer ELI format.

[51] The format is based on using templates to indicate the alphanumeric characters of articles and clauses. The template format can be represented as: https://eur-lex.europa.eu/eli/reg/YEAR/NUMBER/ARTICLE/PARA/POINT/oj

[52] The information is based on a private communication between the author of the thesis with members of the EU Publications Office. The prescribed IRI, while not officially published or documented, currently resolves to the web-page of the legislation, which in this case is the GDPR.

One of the aims of DPVCG is to provide a glossary of concepts associated with legal personal data handling. For this, concepts of DPV itself are considered a glossary of terms, though not explicitly defined as such within the ontology. This use of 'glossary' refers to providing concepts to an adopter in order to represent required information in an interoperable manner. Since these terms do not necessarily arise from a particular legislation, their sources are based on their use as a commonly understood concept or notion within legal domain.

In terms of coverage, GDPRtEXT focuses on terms directly obtained from text of GDPR while DPV focuses on modelling of concepts based on their relevance to defining personal data handling in the legal domain. Therefore, while there is a small overlap in concepts directly associated with GDPR, the two vocabularies differ in provision and definition of terms. For example, GDPRtEXT defines `DataSecurity` based on Article 28 and 32 of GDPR as obligations for Controllers and Processors, while DPV defines `Security` as a purpose. To further exemplify this distinction, DPV aims to offer terms that reflect real-world practices while GDPRtEXT focuses on legal text of GDPR. This is evident from hierarchical taxonomy of concepts in DPV such as for personal data categories and purposes so as to enable modelling of practices across a broader spectrum of use-cases.

5.5.3.2 Representing information about activities

DPV does not provide representation of activities through which it can be compared with GDPRov. Instead, DPV provides concepts to represent 'processes' taking place within an organisation, such as those for ensuring technical and organisational measures, purposes, and processing. The modelling of an instance of personal data handling, which consists of specifying purpose, processing, and technical & organisational measures - can be compared with modelling of plans in GDPRov where a purpose represents a plan and its steps represent processing activities - which can be further annotated with measures and legal bases. Through this, it can be summarised that the focus of GDPRov is on representing activities with granularity in terms of their composition and dependency, and that of DPV is on providing metadata as an overview of processing and data handling practices. With this, it is possible to define an instance of personal data handling using DPV to indicate a high-level summary and expand it using GDPRov to represent details of processes and capture their provenance in ex-ante/ex-post phases.

5.5.3.3 Representing information about consent

The DPV and GConsent both provide concepts to represent information about consent based on requirements specified by GDPR. In this, both share an aim to document context of consent with a view towards establishing its validity and compliance. The difference between the two is based on granularity and use of existing vocabularies to represent this information.

The DPV utilises the model provided by its core or base vocabulary to represent information by specifying consent as a legal basis used to justify processing of personal data. In addition to this, it provides properties to indicate the specific notice displayed to obtain consent, its expiry, obtaining or provision of consent, and its withdrawal. In this, attributes such as timestamp and method used to carry out provision and withdrawal are used to indicate

information regarding how consent was obtained. This is based on updating concept from Consent Receipt with requirements of GDPR.

The DPV does not prescribe or utilise any existing vocabulary to specify information. In contrast, GConsent provides similar concepts as DPV base vocabulary and uses vocabularies of PROV-O and GDPRov to represent information about activities. Given that GConsent was an input to DPV and by extension influential in its modelling of consent, there is a degree of compatibility between the two based on similarity of concepts. In this, GConsent provides a more detailed vocabulary for consent while DPV provides a minimal set of concepts regarding consent but is more expressive in representing information through its taxonomies.

5.5.4 Comparing DPV with SotA

This sections compares the DPV with approaches presented in the SotA in Chapter 3. The aim of this exercise to demonstrate extent of DPV's contributions and present its comparison with approaches in SotA. Since DPV is not presented as a direct contribution of this thesis, its formal evaluation is not within scope of this thesis.

The aim of DPV as established by DPVCG is providing a vocabulary for personal data handling which concerns representation of information relevant for legal compliance - in this case associated with GDPR. Based on this, DPV is a vocabulary useful towards representing information about processing of personal data rather than a framework or methodology that can be used to evaluate compliance. Currently, DPV is not accompanied by any documentation demonstrating its use or application in use-cases, though such activities are planned in near future.

Comparing DPV with other vocabularies in SotA as presented in Chapter 3, DPV provides a large amount of concepts in its top-down taxonomies which can be expanded with additional concepts. This enables it to be adapted and expanded for use-cases. This aspect of the DPV is novel within SotA as no other approach aims to provide a similar taxonomy of concepts, and does not incorporate requirements of extending it for a given use-case such as through sub-classing mechanism.

The DPV base vocabulary provides a compact structure representing personal data handling which aims to represent all relevant information required to evaluate and demonstrate compliance. In this, it bears resemblance to SPECIAL core vocabulary [95] which is self-explanatory given that DPVCG was an extension of SPECIAL's work on its vocabularies. This approach is more suitable for representation or documentation of information from a compliance perspective, and is not intended to be specific to any particular law - though the GDPR clearly has a significant influence on its vocabulary. This is again in contrast to approaches in SotA which are often intended to be applied to a particular legislation and a specific use-case.

The representation of technical and organisational measures is the most distinctive feature of DPV, as currently no other approach within SotA provides a comparable representation of these. While there have been efforts to establish vocabularies regarding specification and representation of privacy policies, these tend to focus on use of concepts such as purpose, processing, data storage, data sharing, third parties - which have been utilised quite commonly in SotA.

SUMMARY

This chapter presented the ontologies created to fulfil research objective $RO3$ along with the methodology used for their development and evaluation. It also provides a comparison of ontologies with related approaches in SotA as presented in Chapter 3.

The ontology engineering process presented in Section 1.3.3 described the methodology used for creating ontologies based on best-practices and guidelines advocated by semantic web community. This included ensuring ontology quality, documentation, and releasing developed resources under an open and permissive license. It also described utilisation of compliance questions presented in Section 4.2 as competency questions.

The presented ontologies consisted of GDPRtEXT, GDPRov, and GConsent. GDPRtEXT, presented in Section 5.2, provides a linked data version of text of GDPR created by extending the official ELI [42] ontology to represent GDPR in a granular manner. It also presents a SKOS glossary of concepts associated with GDPR compliance derived from its text. GDPRtEXT enables association and linking of information with concepts and clauses of GDPR through use of persistent IRIs. It thus fulfils research objective $RO3(a)$. GDPRtEXT extends state of the art by providing an ELI-compatible extension capable of representing GDPR clauses at a granular level. It is also novel in providing a glossary of concepts associated with GDPR compliance.

GDPRov provides an OWL2 ontology for representing activities associated with personal data and consent in ex-ante and ex-post phases. It is presented in Section 5.3. The ontology is based on extending existing ontologies of PROV-O [47] and P-Plan [48] with GDPR terminology to represent activities in context of compliance requirements. GDPRov is novel in use of PROV-O and P-Plan together based on a scientific workflow model to represent processes for GDPR compliance. It provides an extensible model that can be used for representing processes related to compliance which can be extended for representing related processes such as compliance activities and organisational processes. Use of GDPRov enables capturing provenance of plans or snapshots of a system at a given time and document them as evidence of planned and maintained compliance. Use of P-Plan with PROV-O enables associating execution of activities with their intended planning and thereby provides systematic linking of ex-ante and ex-post compliance information. GDPRov thus fulfils research objective $RO3(b)$.

The representation of consent is provided by GConsent - an OWL2 ontology presented in Section 5.4 that fulfils research objective $RO3(c)$. GConsent provides representation of information relevant for evaluation of consent under GDPR obligations and requirements. It extends representation of consent as an artefact in GDPRov and models life-cycle of consent based on the concept of states. State or status of consent reflects its suitability for use as a legal basis for processing and is modelled based on management of consent information from an organisation's perspective. In this GConsent is novel within SotA. GConsent is also novel in provision of concepts related to consent for GDPR as it provides a more detailed and comprehensive vocabulary for representing information regarding consent.

The chapter also presented the DPV vocabulary published by W3C Data Privacy and Vocabularies Community Group (DPVCG). The author of the thesis was an active contributor to DPV, and subsequently DPV shares its aim and bears similarity to research presented

in this thesis. Section 5.5 provides a summary of DPV and compares it with ontologies presented as contributions in this thesis - namely GDPRtEXT, GDPRov, and GConsent. A comparison of DPV with SotA is also provided. DPV is intended to provide a vocabulary for representing information about personal data handling and is not limited to GDPR though it is influenced by it. It reflects a community consensus and is intended to be standardised, thereby providing a strong basis for its adoption.

Through these developed ontologies and the significance of DPV, the primary motivation guiding this research as outlined at the beginning of this thesis (Section 1.1) regarding representing information and associating it with GDPR has been addressed. In the next chapter, use of semantic web technologies to query this information to answer compliance questions, such as those presented in Section 4.2, is presented. The next chapter also presents use of semantic web technologies in validating information to ensure its correctness for assessing GDPR compliance based on constraints presented in Section 4.2.3. The querying and validation of information are minor contributions of this thesis, and satisfy research objectives $RO4$ and $RO5$ respectively. For this, the next chapter utilises real-world use-cases to demonstrate application of developed ontologies of GDPRtEXT, GDPRov, and GConsent to represent, query, and validate information for GDPR compliance.

6 | QUERYING AND VALIDATING INFORMATION FOR GDPR COMPLIANCE

This chapter presents an application of semantic web technologies to query and validate information for GDPR compliance. In this, information is represented using developed vocabularies of GDPRtEXT, GDPRov, and GConsent - as presented in Chapter 5. The queries are represented using SPARQL - a W3C standard for querying RDF - and are based on compliance questions presented in Section 4.2. Validation is carried out based on identified assumptions and constraints presented in Section 4.2.3 and is expressed using SHACL - a W3C standard for representing constraints. The presented work represents minor contributions of this thesis, and fulfils research objective $RO4$ regarding querying of information and $RO5$ regarding validation of information for GDPR compliance.

Section 6.1 presents use of SPARQL to query information for answering compliance questions. Use of SHACL to validate information for GDPR compliance is presented in Section 6.2. The chapter ends with conclusions drawn from this research in Table 6.2.6 regarding novelty of contributions.

6.1 QUERYING INFORMATION USING SPARQL

This section presents creation and utilisation of SPARQL queries to retrieve information relevant for GDPR compliance. Creation of queries is dependant on ontological representation of information being retrieved, which in this case includes use of GDPRov and GDPRtEXT ontologies. As no consent instances needed to be represented, GConsent was not used. This is further explained in Section 6.1.1.

A GDPR preparation guide published by Irish Data Protection Commission was used as source of questions for which corresponding the SPARQL queries were created. The methodology used for this is presented in Section 6.1.2 with a demonstration of developed queries presented in Section 6.1.3. A note on evaluation of this work is presented in Section 6.1.4.

6.1.1 SPARQL queries and ontological representation of information

The research regarding querying presented here is based on the task of retrieving information for answering questions relevant to assessment of compliance. It represents utilisation of technical solutions to automate information retrieval and requires machine-readable data (or metadata). For this, developed ontologies provide concepts and relationships necessary

to express information using GDPR terminology and enable association of information with clauses and concepts of GDPR.

The compliance questions, as presented in Section 4.2, do not use a specific ontology or vocabulary but instead are based in natural language and use legal terminology. In order to utilise technological solutions for answering them, it is necessary to first convert these questions into queries using ontological concepts. As research presented in this thesis derives motivation for use of semantic web technologies which use RDF for representing information, querying this utilises SPARQL to retrieve this information.

Ontologies used in SPARQL queries must match ontologies used in representation of information it aims to retrieve. Differences in ontologies hamper effective execution of queries with potential returning of invalid or empty results. Creation of these SPARQL queries is therefore specific to ontologies of GDPRov and GDPRtEXT used for information representation.

6.1.2 Methodology

The methodology used for creation of SPARQL queries is based on utilisation of GDPRov to represent concepts and GDPRtEXT to link information to GDPR. SPARQL queries thus created aim to retrieve information relevant to answering a question rather than show evaluation or assessment of compliance. While compliance questions presented in Section 4.2 provide a basis for construction of semantic queries using SPARQL, presented application of SPARQL utilises a real-world use-case of questions to provide an demonstration of this research.

6.1.2.1 Utilising compliance questions from GDPR readiness guide published by DPC

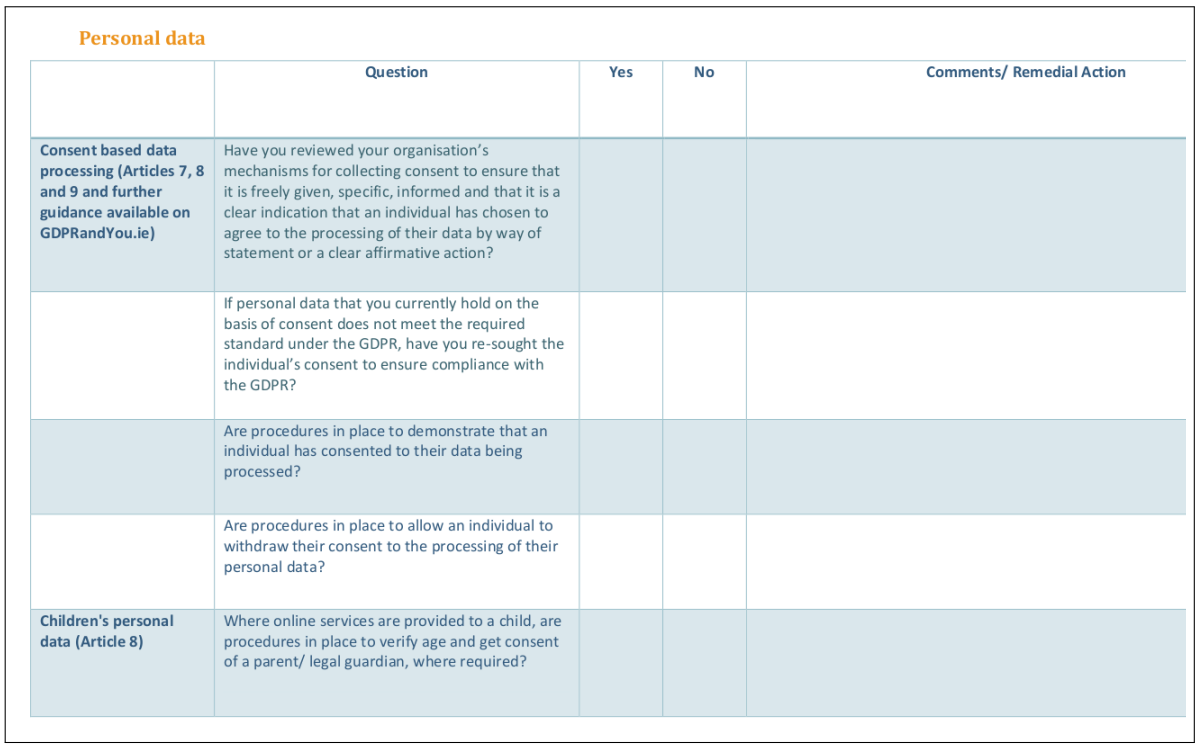
The application of SPARQL utilised the guide titled "Preparing Your Organisation for the GDPR – A Guide for SMEs" published by Data Protection Commission of Ireland (DPC) as basis for (compliance related) questions which were represented using SPARQL as semantic queries. The guide was published by DPC in 2017 to help organisations in assessing their readiness towards GDPR compliance requirements. It is accessible online[1] and consists of a 'table' (see Figure 6.1) containing questions regarding information about an organisations processing activities. The guide was chosen based on its simplicity in terms of questions, its intended use in evaluating information associated with compliance, and locality of Irish DPC with respect to the author.

The guide divides questions into contextual sections based on addressing specific GDPR articles and obligations.

6.1.2.2 Steps of the methodology

The steps followed in utilising questions in the guide to create SPARQL queries and demonstrate their application were as follows:

[1] http://gdprandyou.ie/wp-content/uploads/2017/12/A-Guide-to-help-SMEs-Prepare-for-the-GDPR.pdf



Personal data

	Question	Yes	No	Comments/ Remedial Action
Consent based data processing (Articles 7, 8 and 9 and further guidance available on GDPRandYou.ie)	Have you reviewed your organisation's mechanisms for collecting consent to ensure that it is freely given, specific, informed and that it is a clear indication that an individual has chosen to agree to the processing of their data by way of statement or a clear affirmative action?			
	If personal data that you currently hold on the basis of consent does not meet the required standard under the GDPR, have you re-sought the individual's consent to ensure compliance with the GDPR?			
	Are procedures in place to demonstrate that an individual has consented to their data being processed?			
	Are procedures in place to allow an individual to withdraw their consent to the processing of their personal data?			
Children's personal data (Article 8)	Where online services are provided to a child, are procedures in place to verify age and get consent of a parent/ legal guardian, where required?			

Figure 6.1: Questions for information required to assess compliance - Page 10 of "Preparing Your Organisation for the GDPR - A Guide for SMEs" published by Ireland's Data Protection Commission

1. Analyse questions within the document to identify corresponding concepts and relationships in GDPRov and GDPRtEXT (see below). The questions largely concerned details of processing activities and organisational practices and therefore did not require use of GConsent.
2. Represent questions as SPARQL queries using GDPRov and GDPRtEXT (see below)
3. Create a synthetic use-case based on processing of personal data with GDPRov and GDPRtEXT used to represent information (see Section 6.1.3)
4. Execute SPARQL queries over use-case to retrieve answers for compliance questions (see Section 6.1.3)
5. Evaluate queries based on subjective criteria of - a) Extent of answering compliance questions b) Suitability of retrieved results in answering compliance questions (see Section 6.1.4)

6.1.2.3 Analysis of GDPR Readiness Guide

The guide contains 63 questions across 13 pages that are presented in 9 sections. Its analysis consisted of categorising questions based on requirements of information, relation to phases of compliance, and whether they were suitable to be implemented as SPARQL queries. The analysis was recorded and published online[2] as a spreadsheet with comments describing interpretation of each question's information requirements.

The first set of questions on page 1 concern consent and personal data and are structurally different than other sets in that they are more abstract and generic and concern over-

[2] https://w3id.org/GDPRep/checklist-demo/notes

all practices concerning processing of personal data by an organisation. These questions are described under 'general' category with other groups of questions having their category mentioned explicitly within the document. Questions in general category require information and practices associated with consent and personal data. Other categories contain questions which enquire explicitly about activities and mechanisms regarding compliance to specific obligations.

The questions were analysed and categorised based on their intended requirements towards information required for compliance. The three categories identified through this exercise were - demonstrative, evaluative, and assistive - based on requirements of information associated with them. Demonstrative questions require answers that satisfy the question and do not need further actions or processing based on information. Assistive questions provide information that needs to be directly evaluated for compliance, with 'assistive' indicating information that assists evaluation of compliance. Evaluative questions retrieve information whose evaluation requires further information retrieved through additional questions based on provided information. The primary difference assistive and evaluative questions is whether they retrieve information which can be evaluated as is for compliance or whether it requires additional questions to retrieve further information. These terms used for categorisation do not relate to any specific methodology used in legal compliance, but are useful to analyse questions from an information management perspective.

Questions were also analysed based on whether they relate to or require information regarding activities in ex-ante and ex-post phases. The questions do not explicitly provide an indication of whether they enquire about a model of processing (ex-ante) or logs (ex-post). The distinction was made based on whether a question concerned information about practices, plans, or intentions regarding processing of personal data - in which case it was deemed to enquire about ex-ante information. Similarly, if a question concerned past execution of activities or records of activities - it was specified to enquire about ex-post information. In some cases, questions were specified to enquire about both ex-ante and ex-post information based on potential application in both phases.

An overview of the questions is provided in Table 6.1. It assigns an ID for each question to enable associating it with corresponding SPARQL queries and for linking related questions in analysis. The column '*Category*' reflects category of question mentioned within the guide, with 'general' used for initial generic questions. '*Title*' refers to title of text within the guide, and column '*GDPR*' refers to an explicit mention of a GDPR clause within the question or its description.

Table 6.1: Questions provided in the GDPR Readiness Guide

ID	Category	Title	GDPR
G1	General	Categories of personal data and data subjects	
G2	General	Elements of personal data included within each data category	
G3	General	Source of the personal data	
G4	General	Purposes for which personal data is processed	
G5	General	Legal basis for each processing purpose (non-special categories of personal data)	

(Cont'd on following page)

Questions provided in the GDPR Readiness Guide (cont'd)

ID	Category	Title	GDPR
G6	General	Special categories of personal data	
G7	General	Legal basis for processing special categories of personal data	
G8	General	Retention period	
G9	General	Action required to be GDPR compliant?	
P1	PersonalData	Validity of Consent	7,8,9
P2	PersonalData	Retrospective Consent	7,8,9
P3	PersonalData	Demonstration of Consent	7,8,9
P4	PersonalData	Withdraw consent for processing	7.8.9
P5	PersonalData	Children's Personal Data	8
P6	PersonalData	Legitimate interest based data processing	
R1	Rights	Subject Access Requests (SARs)	15
R2	Rights	Subject Access Requests (SARs) Response Time	15
R3	Rights	Data Portability	20
R4	Rights	Deletion and Rectification	16,17
R5	Rights	Right to restriction of processing	18
R6	Rights	Right to object to processing	21
R7	Rights	Halt processing after right to object	21
R8	Rights	Profiling and automated processing	22
R9	Rights	Right to obtain human intervention	22
R10	Rights	Restrictions to data subject rights	23
A1	AccuracyRetention	Purpose Limitation	
A2	AccuracyRetention	Data minimisation	
A3	AccuracyRetention	Accuracy	
A4	AccuracyRetention	Retention	
A5	AccuracyRetention	Retention Legal Obligations	
A6	AccuracyRetention	Destroy data securely	
A7	AccuracyRetention	Duplication of records	
T1	Transparency	Transparency to customers and employees	12,13,14
T2	Transparency	Provide Information listed in Article 13	13
T3	Transparency	Provide Information listed in Article 14	14
T4	Transparency	Provide information when engaging	
T5	Transparency	Provide information on facilitating rights	
C1	ControllerObligations	Supplier Agreements	27,28,29
C2	ControllerObligations	Data Protection Officers	37,38,39
C3	ControllerObligations	Reasons for not having a DPO	37,38,39
C4	ControllerObligations	Escalation procedures	37,38,39
C5	ControllerObligations	Escalation procedures through a DPO	37,38,39
C6	ControllerObligations	Data Protection Impact Assessments (DPIAs)	35
S1	DataSecurity	Risks involved in processing data	32
S2	DataSecurity	Documented Security Program	32
S3	DataSecurity	Resolving security related issues	32
S4	DataSecurity	Designated individual for security	32
S5	DataSecurity	Encryption	32
S6	DataSecurity	Removing information	32
S7	DataSecurity	Restoring access	32
B1	DataBreach	Documented incident plans	33,34
B2	DataBreach	Regular reviews	33,34
B3	DataBreach	Notifying authorities	33,34

(Cont'd on following page)

Questions provided in the GDPR Readiness Guide (cont'd)

ID	Category	Title	GDPR
B4	DataBreach	Notifying data subjects	33,34
B5	DataBreach	Documentation of data breaches	33,34
B6	DataBreach	Co-operation procedures for data breach	33,34
I1	InternationalDataTransfer	Data transfer outside EEA	44,45,46,47,48,49,50
I2	InternationalDataTransfer	Special category of Personal Data in Transfer	44,45,46,47,48,49,50
I3	InternationalDataTransfer	Purpose of Transfer	44,45,46,47,48,49,50
I4	InternationalDataTransfer	Transfer Recipients	44,45,46,47,48,49,50
I5	InternationalDataTransfer	Transfer Details	44,45,46,47,48,49,50
I6	InternationalDataTransfer	Legality of international transfers	
I7	InternationalDataTransfer	Transparency	

Table 6.2 presents a summarised view of analysis of questions presented in Table 6.1. The complete information along with additional comments and fields is available in an online version of analysis. In the table, column '*Type*' indicates type of query based on categorisation as demonstrative, assistive, evaluative based on description in previous sections. Column '*Data*' provides information on information required for the question, including results of other queries. Relation of question to ex-ante phase of compliance is reflected by column '*E/A*' and ex-post phase by column '*E/P*' using boolean `Y/N` values. Column '*SPARQL*' indicates whether a SPARQL query was constructed for the corresponding question, with `N` indicating that a query was not constructed. Column '*GDPRov*' indicates whether the latest iteration (v0.7) of GDPRov provides concepts and relationships to answer the question, with a value of `Y` indicating that it does, `N` indicating it does not provide concept, and `S` indicating information to be out of scope. Where a query was not constructed, the reason can be inferred by combining values in *SPARQL* and *GDPRov* columns - for example, where concepts were out of scope for GDPRov query was not constructed due to lack of concepts.

Table 6.2: Analysis of compliance questions specified in Table 6.1

ID	Type	Data	E/A	E/P	SPARQL	GDPRov
G1	Demonstrative	personal data, data subjects	Y	N	Y	Y
G2	Demonstrative	personal data	Y	N	Y	Y
G3	Demonstrative	personal data, steps that collect data, entities that provide data	Y	Y	Y	Y
G4	Demonstrative	results of G1, processes acting on data	Y	N	Y	Y
G5	Demonstrative	results of G4, processes acting on data	Y	N	Y	Y
G6	Demonstrative	special category personal data	Y	N	Y	Y
G7	Demonstrative	results of G6, steps that collect data, steps that store data	Y	N	Y	Y
G8	Not-Implemented	results of G1, steps that store data			N	N
G9	Not-Implemented				N	S
P1	Assistive	consent, steps that acquire consent	Y	N	Y	Y
P2	Not-Implemented				N	S
P3	Evaluative	consent	Y	Y	Y	Y

(Cont'd on following page)

Analysis of compliance questions specified in Table 6.1 (cont'd)

ID	Type	Data	E/A	E/P	SPARQL	GDPRov
P4	Evaluative	steps that withdraw consent	Y	N	Y	Y
P5	Evaluative	steps that acquire consent, steps for age verification	Y	N	Y	Y
P6	Assistive	steps that process personal data	Y	N	Y	Y
R1	Assistive	steps that handle SAR	Y	N	Y	Y
R10	Not-Implemented				N	S
R2	Assistive	steps that handle SAR	N	Y	N	Y
R3	Evaluative	steps that address right to data portability	Y	N	Y	Y
R4	Evaluative	steps that address right to rectification	Y	N	Y	Y
R5	Assistive	data subject request, steps that process personal data	N	Y	N	Y
R6	Not-Implemented				N	Y
R7	Evaluative	steps that process personal data	Y	N	Y	Y
R8	Assistive	steps that make automated decisions, consent	Y	Y	Y	Y
R9	Assistive	steps that make automated decisions, right to contest automated decisions	Y	N	Y	Y
A1	Evaluative	personal data, consent, steps that involve personal data through use, share, store	Y	Y	Y	Y
A2	Assistive	personal data, steps that process personal data	Y	Y	Y	Y
A3	Not-Implemented				N	S
A4	Not-Implemented				N	S
A5	Not-Implemented				N	S
A6	Assistive	steps that delete data	Y	N	Y	Y
A7	Not-Implemented				N	S
T1	Not-Implemented				N	S
T2	Assistive	steps that collect personal data	Y	N	Y	Y
T3	Assistive	steps that collect personal data	Y	N	Y	Y
T4	Not-Implemented				N	S
T5	Not-Implemented				N	S
C1	Not-Implemented				N	S
C2	Not-Implemented				N	Y
C3	Not-Implemented				N	S
C4	Not-Implemented				N	S
C5	Not-Implemented				N	S
C6	Assistive	steps part of the DPIA process	Y	N	Y	Y
S1	Assistive	steps that process data	Y	N	Y	Y
S2	Not-Implemented				N	S
S3	Not-Implemented				N	S
S4	Not-Implemented				N	S
S5	Not-Implemented	steps that share data			N	N
S6	Not-Implemented				N	Y
S7	Not-Implemented				N	N
B1	Evaluative	processes or plan that address security incidents	Y	N	Y	Y
B2	Not-Implemented				N	S
B3	Evaluative	processes or plans for notifying DPC	Y		Y	Y

(Cont'd on following page)

Analysis of compliance questions specified in Table 6.1 (cont'd)

ID	Type	Data	E/A	E/P	SPARQL	GDPRov
B4	Evaluative	processes or plans for notifying data subjects of a data breach			Y	Y
B5	Not-Implemented				N	Y
B6	Not-Implemented				N	S
I1	Evaluative	steps that share data	Y	Y	Y	Y
I2	Evaluative	results from I1, category of personal data	Y	N	Y	Y
I3	Assistive	steps that share data	Y	Y	Y	Y
I4	Evaluative	steps that share data	Y	Y	Y	Y
I5	Not-Implemented				N	Y
I6	Not-Implemented				N	Y
I7	Not-Implemented	steps that share data			N	S

Information regarding GDPRov is also provided since creation of SPARQL queries from GDPR readiness guide was carried out in earlier stages of GDPRov's iterations and before enforcement of GDPR in May 2018. Therefore, some questions were deemed to be ambiguous or lacking legal information on information necessary for compliance. The queries and constraints presented in Section 6.2 were developed at a later stage when GDPR had seen significant attention and interpretation and present a more mature implementation.

6.1.2.4 Creation of SPARQL queries

Creation of SPARQL queries involved analysis of text of a question to identify relevant concepts and relationships in GDPRov useful towards expressing the question as a semantic query in SPARQL as well as representing information required to answer the question. In this, some questions were found to be subjective or qualitative based on information they required and thus could not be expressed as SPARQL queries. For example, *Question C3* is about reasons for not having a DPO. Such questions are indicated as not implemented in Table 6.2.

A total of 33 SPARQL queries were created based on analysis of compliance questions and their requirements. The queries utilised GDPRov and GDPRtEXT ontologies to specify information associated with questions. The SPARQL queries were published online[3] with separate files for each query associated with a question, and a common file containing common prefixes used in all queries.

As an example, Listing 9 contains corresponding SPARQL query for question `G5` which concerns legal basis used to justify processing of personal data. The query retrieves information about steps and processes along with legal basis for their operation in ex-ante phase using GDPRov. Within this, the query specifically retrieves steps which are defined as being part of a process and use some form of personal data, where the legal bases can be associated with individual steps or with a process.

[3] https://w3id.org/GDPRep/checklist-demo/sparql-queries

```
PREFIX rdfs:     <http://www.w3.org/2000/01/rdf-schema#>
PREFIX gdprov:   <http://purl.org/adaptcentre/openscience/ontologies/gdprov#>
PREFIX gdprtext: <http://purl.org/adaptcentre/openscience/ontologies/GDPRtEXT#>

SELECT DISTINCT ?process ?legal WHERE {
  ?data a ?data_type .
  ?data_type rdfs:subClassOf gdprov:PersonalData .
  ?step a ?step_type .
  ?step_type rdfs:subClassOf gdprov:DataStep .
  ?step gdprov:usesData ?data .
  ?step gdprov:isPartOfProcess ?process .
{
  OPTIONAL { ?step gdprov:hasLegalBasis ?legal } .
} UNION
{
  OPTIONAL { ?process gdprov:hasLegalBasis ?legal } .
}
} ORDER BY ?process
```

Listing 9: SPARQL query representing compliance question `G5` concerning legal basis for processing

6.1.3 Demonstration using synthetic use-case

To demonstrate application of queries, a synthetic use-case was created using GDPRov and GDPRtEXT to represent information. The use-case is based on the scenario of an online shopping service that allows users to order products. RDF representations of processes and personal data associated with the use-case were created and queried using created SPARQL queries to retrieve information regarding compliance. The implementation was published online[4] along with its data and code in a public repository[5]. The use-case of an online shopping service is based on its prevalence in real-world and provides a sufficient representation of purposes, legal bases, processing operations, and third parties involved. The use-case is intended to provide information for SPARQL to query and as such its complexity does not have a significant bearing on design of queries as long as queried concepts have been represented.

6.1.3.1 Use-case: Online shopping service that shows ads

Within the use-case, users can shop for products using an online service i.e. a website. Users have an option to establish an account to receive discounts and special offers for products offered. Ads are served to users and are generated by a Third Party. The sign-up process collects personal data such as name, address, email, and contact number. While ordering products, users are requested to provide sensitive information for transactions about their bank account or credit cards.

Personal data is represented by sub-classing `gdprov:PersonalData` as `CustomerInfo` in the use-case's namespace for representing information about users. Similarly `SensitiveData` is sub-classed as `gdprov:BankingInfo` for representing banking and financial information. Processes for handling obligations and rights are expressed using GDPRov.

[4] https://w3id.org/GDPRep/checklist-demo
[5] http://openscience.adaptcentre.ie/GDPR-checklist-demo/demo/

The sign-up process enables an user to provide information which is used for personalisation and ads and collects user's consent. As a final step, Fact++[6] semantic reasoner was used to derive additional facts and to ensure logical consistency of information.

6.1.3.2 Implementation

The online demo provides an execution of created SPARQL queries over data defined for the use-case. This represents automation of answering compliance questions using retrieved information. The demo is intended to showcase how a static GDPR readiness checklist or questionnaire can be made more interactive and automated using semantic web technologies.

The demo is provided as a single web page application with questions from GDPR readiness checklist provided in their natural language form and ID followed by its corresponding SPARQL query. The results for each query are retrieved on page load from a SPARQL endpoint[7] containing RDF data about the use-case. The demo uses tools YASQE[8] to present SPARQL queries with syntax highlighting and YASR[9] to represent results of queries in an interactive fashion.

The results of each query contain information associated with answering relevant questions. SPARQL query regarding question `G5` is presented in Listing 9 which enquires about legal obligations and whose results express steps and processes along with their legal obligations. The query and results as presented in the demo are depicted in Figure 6.2. In this, the results consist of five rows - of which three are processes that handle various rights and therefore are not accompanied with any legal basis[10]. The remaining two results represent processes associated with provision of service, of which *OrderProcess* represents 'ordering a product' and uses legitimate interest as its legal basis, and *NewUserSignUpProcess* collects information about an user and uses legal basis of given consent.

6.1.4 Evaluation

The aim of this work was to represent compliance questions using SPARQL in order to retrieve information represented in RDF regarding processing of personal data. The demonstration using a synthetic use-case provided basis for exploring the application of created SPARQL queries by using GDPRov and GDPRtEXT for ontological representations of data. The evaluation of this work, while not being exhaustive, demonstrates creation of SPARQL queries and their application over a given use-case.

In terms of coverage of compliance questions represented as SPARQL queries, the exercise could not represent all questions within the GDPR readiness guide. Reasons include a lack of knowledge regarding representation of ambiguous information such as 'indefinite' storage periods and their legal validity, and a query being out of scope for the research question of this thesis. Table 6.2 presents an indication of these through *SPARQL* and *GDPRov* columns.

[6] http://owl.cs.manchester.ac.uk/tools/fact/
[7] http://openscience.adaptcentre.ie/sparql
[8] http://yasqe.yasgui.org/
[9] http://yasr.yasgui.org/
[10] The processes handling rights should utilise the legal basis of requirements specified by law since GDPR requires the provision of rights.

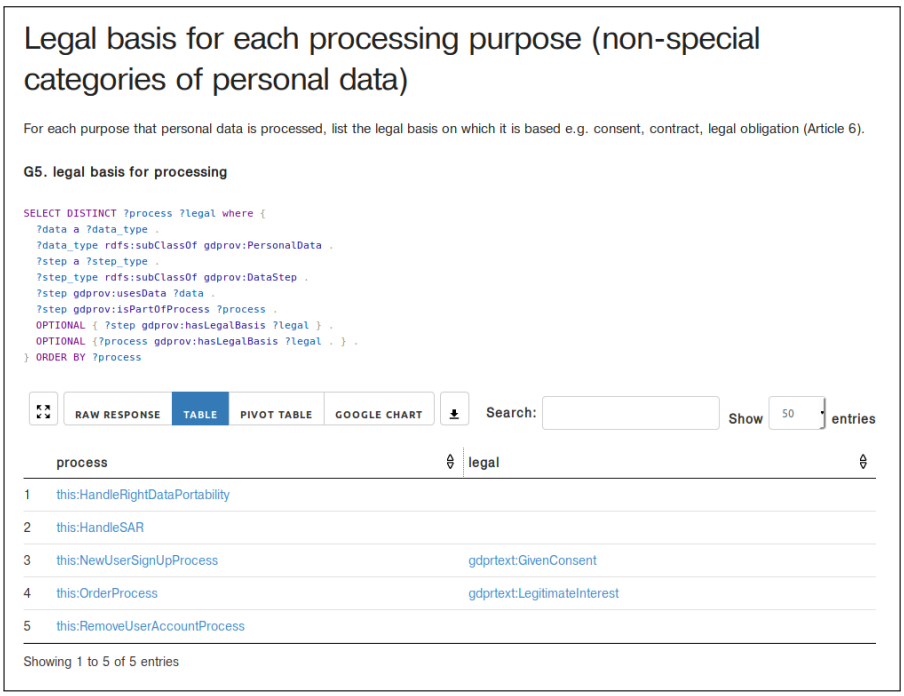


Figure 6.2: Retrieving information using SPARQL for query G5 in GDPR readiness checklist

Since the goal of this exercise was demonstrating how questions related to GDPR compliance could be expressed in SPARQL, its evaluation consisted of determining the extent to which this was possible. The expression of compliance questions using SPARQL is not novel in itself as approaches in SotA present their use of SPARQL in querying information related to GDPR compliance - such as in SPECIAL, MIREL, and DAPRECO projects. However, details of their creation and implementation are sparse as pointed out by the analysis in Section 3.7, which makes it difficult to compare application of SPARQL queries for retrieving information associated with GDPR compliance as presented here.

Of the total 63 questions within the GDPR readiness guide, 32 questions have corresponding SPARQL queries created and used in the demo. Of 31 questions that were not implemented, 20 questions were considered out of scope as they do not relate to the research question, with the other 3 questions lacking corresponding concepts in GDPRov to create SPARQL queries. Of these, question `G8` concerning retention periods for personal data can be expressed using Time ontology [184]. The other two, `S5` and `S7` require specification of information associated with information management and governance procedures utilised within an organisation. While these are technically not outside the scope of GDPRov, they require a larger understanding of how such processes are specified and managed and commonly involve use of specifications to denote practices - for example ISO/IEC 27001 describing a framework for information management and protection. Some questions not considered within scope concern information not associated with processing of personal data or consent, but which can be represented as activities using GDPRov. These include questions `C1` concerning agreements between entities or question `C4` concerning escalation procedures involving DPO.

The application of SPARQL for querying information associated with GDPR compliance was published in a peer-reviewed publication [57] at SEMANTiCS conference - which provided its exposure to an audience of industry and academic participants. The publication

has received 6 citations to date (excluding self-publications), which includes one approach [183] which utilises modelling of concepts using GDPRov towards annotating DFDs (data flow diagrams) with information for analysing compliance, and provides an example of a SPARQL query to retrieve information about the data flows.

6.2 Validating Information using SHACL

This section presents application of SHACL to validate information based on requirements of GDPR compliance. SHACL is utilised as a validation mechanism to create a test-driven approach where information regarding processing activities is represented using ontologies and is first checked for correctness and then for compliance. In this, constraints presented in Section 4.2.3 are utilised to determine correctness and compliance of information, and questions presented in Section 4.2 are used to retrieve information for compliance. Both constraints and queries are linked to GDPR using GDPRtEXT.

SPARQL is intended to express queries that retrieve information while SHACL is intended to express validation via constraints. SPARQL queries can be utilised as a validation mechanism by retrieving information violating the constraint. However, SHACL provides additional features and capabilities such as persistence of results, recursive constraints, modular composition of constraints, and more importantly - the ability to customise information within constraint and results - which is used here to enable linking of validations and results to relevant clauses of GDPR using GDPRtEXT.

The results of SHACL validations are persisted to create a 'compliance graph' which enables querying for information regarding compliance, and provides more efficient testing based on reuse of ex-ante test results in ex-post testing. The approach is demonstrated using a proof-of-concept implementation based on evaluation of consent information on a real-world website and using GDPRov, GConsent, and GDPRtEXT to represent information. The approach and implementation have been published in peer-reviewed publications concerning the conceptual model of testing approach [59], construction of a knowledge graph from information about GDPR compliance [58], and implementation testing compliance of given consent on a real-world website [60]. All resources regarding this work have been published online[11] under an open and permissive license (CC-by-4.0).

The work presented in this section fulfils research objectives $RO5$ and demonstrates the following:

1. Utilises SHACL to validate information for GDPR compliance.
2. Expresses compliance as a test-driven exercise similar to the concept of unit-testing in software engineering.
3. Utilises results of testing ex-ante information for testing of ex-post information in order to reduce the number of tests required.
4. Constructs a compliance graph by storing validation results based on concept of knowledge-graph.
5. Demonstrates use of compliance graph in retrieving and documenting information regarding GDPR compliance.

A description of the approach is provided in Section 6.2.1 which presents role of SHACL

[11] https://w3id.org/GDPRep/semantic-tests

as a validation mechanism and creation of a compliance graph containing information relevant for compliance. Creation of SHACL constraints to represent constraints expressed in natural language in Section 4.2.3 is presented in Section 6.2.2, with an argument for utilisation of ex-ante validations for evaluation of ex-post information presented in Section 6.2.3. A proof-of-concept implementation demonstrating application of approach is presented in Section 6.2.4, with creation of documentation and compliance reports presented in Section 6.2.5.

6.2.1 Validation Model

The validation model represents an abstract and generalised overview of the validation approach. It does not utilise any specific ontology for information representation and can be implemented with any ontologies as presented in SotA or this thesis. In addition, use of SHACL can be substituted with another technology as long as it supports validation of information. The intention of this generalised overview is to present developed work as applicable to a larger set of technologies, with a specific implementation using semantic web ontologies presented in this thesis.

Figure 6.3 presents a visual overview of the approach for validation model. The terminology consists of terms specified by SHACL - which includes *data graph* indicating RDF data to be evaluated using SHACL, *completeness* indicating sufficiency of information i.e. all required data is present, and *validation* as the process of evaluating constraints on data graph. In addition to these, the term *compliance graph* is introduced by this work for indicating a RDF data graph containing information relevant for GDPR compliance. The terms *testing* and *evaluating* act as synonyms to *validation* and are used interchangeably to refer to the same process.

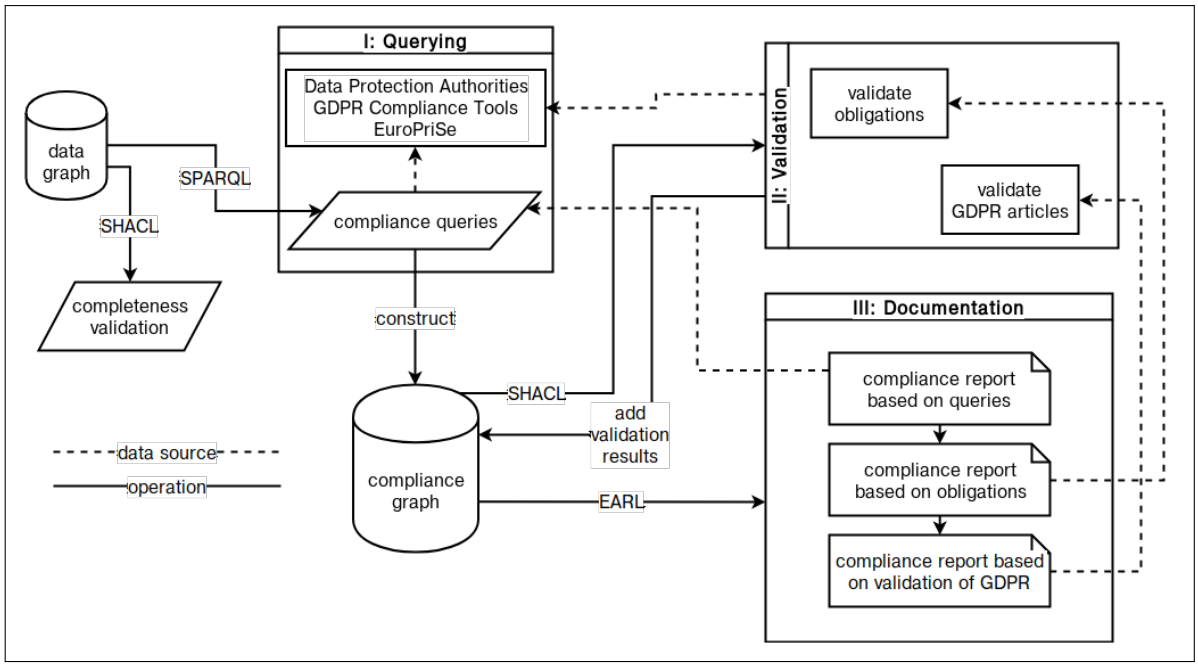


Figure 6.3: Overview of approach utilising SHACL to validate information for GDPR compliance

The approach described in the figure consists of three steps - (i) Querying, (ii) Valida-

tion, and (iii) Documentation. A data graph acts as input and provides information about activities associated with processing of personal data and consent represented in RDF. The data graph needs to be first checked for 'completeness' - i.e. ensuring required information is present before it can be evaluated for GDPR compliance. After this, the first step of querying retrieves information for answering compliance queries by using SPARQL. This information is then added to a 'compliance graph' which is separate from the data graph and stores information for determining and documenting compliance.

In the second step of validation, SHACL constraints representing obligations and requirements of GDPR are executed over compliance graph with results added back to compliance graph. The SHACL constraints and evaluated results are linked to specific GDPR obligations and articles. At this point, compliance graph contains information required to answer compliance questions and an evaluation using SHACL. This information is linked to relevant GDPR clauses. The graph thus enables retrieval of information relevant to compliance based on a specific question, concept, or GDPR clause.

In the third and final step, information within compliance graph is used for documentation of compliance information based on compliance questions, fulfilment of obligations, or coverage of GDPR articles. It queries the compliance graph using SPARQL and retrieves information along with their link or relevance to specific GDPR clauses. The results of these can then be persisted or demonstrated using any presentation medium - such as a webpage, dashboard, or even a data file.

The use of RDF makes SPARQL and SHACL the default choices for querying and validation respectively given their status as standards. However, the model presents a modular approach for querying, validating, and documenting information relevant to compliance. This is to enable use of alternative technologies for carrying out tasks associated in a particular step. For example, ShEx - another validation standard - could be used in lieu of SHACL to express constraints over RDF data.

The steps only represent a separation of concerns within the model. In practical uses, such as one presented in this thesis, the first and second steps are combined to consolidate validations associated with correctness and GDPR obligations. This is based on the assumption that missing information (which is checked by completeness validations) is a failing condition in evaluation of compliance. The constraints presented in Section 4.2.3 thus incorporate expression of validations for both completeness and obligations.

6.2.2 Creation of SHACL constraints

6.2.2.1 Ontologies for expressing SHACL constraints

Section 6.1 mentioned dependency of SPARQL queries on underlying data model which necessitates utilisation of the same ontological representations as those used in information to be queried. The same argument applies for validation of information using SHACL, where constraints must utilise the same ontologies as those used in the RDF it aims to validate. An alternative is using mapping tables to convert ontologies used in data to a common ontology used in validation constraints - however, this will be a difficult, if not impossible, exercise due to complexities of finding a common model in all the ontologies that can be potentially used to represent information, such as those within state of the art.

The constraints presented here use developed ontologies from Chapter 5 as: GDPRov to represent activities associated with processing of personal data and consent, GConsent to represent information about consent relevant for compliance, and GDPRtEXT to link information with concepts and clauses of GDPR. In this, the use of GDPRov and GConsent is complimentary in some constraints given their overlap in representing concepts associated with consent. GDPRtEXT is used to link a constraint to a clause within GDPR to indicate its relevancy regarding compliance. It is also used to link validation results with clauses in GDPR to enable querying of results based on GDPR articles, as shown later in Section 6.2.5.

6.2.2.2 Extending SHACL concepts to associate information with GDPR

In the assessment of information for GDPR compliance, some constraints cannot be evaluated automatically based on their qualitative requirements. For example, constraints associated with given consent that aim to evaluate whether it was 'freely given' or 'unambiguous'. These constraints need to be manually evaluated and their results added to compliance graph. To distinguish such constraints, the SHACL concept of `NodeShape` for representing a shape was extended with a sub-class as `Constraint` with further sub-classes of `ManuallyCheckedConstraint` and `AutomaticallyCheckedConstraint` representing constraints that should be checked manually and automatically respectively. This is presented in Listing 10. The property `linkToGDPR` was created to link information with clauses of GDPR with the range `eli:LegalResourceSubdivision` to enable associating it with any granular part of GDPR - such as a chapter, article, paragraph, or sub-paragraph - based on GDPRtEXT's uses of this concept in representing structure of GDPR.

```
:Constraint rdfs:subClassOf sh:NodeShape ;
    rdfs:label "Constraint" .
:AutomaticallyCheckedConstraint rdfs:subClassOf :Constraint, sh:NodeShape ;
    rdfs:label "Automatically Checked Constraint" .
:ManuallyCheckedConstraint rdfs:subClassOf :Constraint, sh:NodeShape ;
    rdfs:label "Manually Checked Constraint" .

:linkToGDPR a rdfs:Property ;
    rdfs:range eli:LegalResourceSubdivision ;
    rdfs:label "linkToGDPR" .
```

Listing 10: Extending SHACL `NodeShape` to express manual and automated checking of constraints

The constraints utilise both GDPRov and GConsent where appropriate and feasible so as to verify using both ontologies. For example, Listing 11 presents a constraint for checking whether each instance of consent is associated with one and only one Data Subject. In it, the concept of Data Subject could be used from GDPRov or GConsent since they both feature it. Therefore, `sh:or` in SHACL enables representing the condition where either of those could be used to express a Data Subject. The constraint is linked to the Article 4-11 of GDPR using property `linkToGDPR`, and provides a human readable message when it fails using the SHACL property `sh:message`.

```
:ConsentHasDataSubject a sh:PropertyShape, :AutomaticallyCheckedConstraint ;
    sh:name "Consent --> Data Subject" ;
    :linkToGDPR gdpr:article4-11 ;
    sh:path gc:isConsentForDataSubject ;
    sh:minCount 1;
    sh:maxCount 1;
    sh:or ( [ sh:class gc:DataSubject ] [ sh:class gdprov:DataSubject ] ) ;
    sh:message "Consent should be linked to Data Subject" .
```

Listing 11: SHACL constraint checking Data Subject associated with consent

6.2.2.3 Using SHACL-SPARQL

SHACL-SPARQL[12] is an extension of SHACL core features and provides use of SPARQL queries to retrieve information failing the associated constraint. Listing 12 presents alternative representations of the same constraint in SHACL core and SHACL-SPARQL. The constraint aims to ensure all consent instances have a timestamp. The SHACL-SPARQL constraint features a SPARQL query that filters instances that have a specified timestamp based on properties provided by GConsent, GDPRov, or PROV-O, while the SHACL core representation uses a `PropertyShape` to assess the same.

```
# SHACL-SPARQL
sh:select "
    SELECT $this WHERE {
        FILTER NOT EXISTS { $this gc:atTime ?time } .
        FILTER NOT EXISTS { $this prov:generatedAtTime ?time } .
        FILTER NOT EXISTS { $this a gdprov:ConsentAgreementTemplate } .
    } " .
```

```
# SHACL Core
_:ConsentHasTimestamp a sh:PropertyShape ;
    sh:or (
        [ sh:path gc:AtTime . sh:minCount 1; ] ;
        [ sh:path prov:generatedAtTime . sh:minCount 1; ] ;
        [ sh:path gdprov:ConsentAgreementTemplate . sh:minCount 1; ] ;
    ) .
```

Listing 12: Expressing the same constraint in SHACL-SPARQL and in SHACL core

The advantages of using SPARQL queries in SHACL constraints is access to information in instances that fail validation. This is useful in inserting information about the instance in validation results - such as ID or IRI of the node or even a specific triple that needs correction or verification. The use of SHACL-SPARQL also allows use of SPARQL queries from Section 6.1 by modifying them to retrieve information that will fail the constraint.

6.2.2.4 Validating manually evaluated constraints

For manually evaluated constraints represented using `ManuallyCheckedConstaint`, the result arising from its assessment indicates whether the constraint fails or passes, and is therefore a boolean value. Therefore, assessment of manually checked constraints is based

[12] https://www.w3.org/TR/shacl/#shacl-sparql

on verifying a boolean value associated with the constraint through the SHACL property `sh:hasValue` which indicates expected value of a property. An example of this is presented in Listing 13 which represents a constraint checking whether given consent was freely given. The assessment is based on an explicitly added triple within the data graph through the property `m:consentIsFreelyGiven` whose value must be true to indicate a manual inspection of the condition that consent must be freely given. The messages of a manually checked constraint are prefixed with *(MANUAL-TEST)* to indicate their qualitative nature in human-intended messages.

```
:ValidconsentIsFreelyGiven a sh:PropertyShape, :ManuallyCheckedConstraint ;
    :linkToGDPR gdpr:article4-11 ;
    sh:name "Consent == Freely Given" ;
    sh:path m:consentIsFreelyGiven ;
    sh:hasValue true ;
    sh:message "(MANUAL-TEST) Consent should be freely given" .
```

Listing 13: Evaluating manually checked constraints using boolean values

6.2.3 Utilising ex-ante test results for ex-post validations

Based on distinguishing information about activities in ex-ante and ex-post phases, the constraints will also need to be expressed to evaluate these phases separately. This will cause duplication of evaluations based on testing of same information across both phases. For example, in evaluating whether given compliance is compliant with requirements of GDPR compliance - which is an ex-post evaluation of compliance - information about criteria such as whether consent was informed are based on artefacts shown during request for consent. The information shown when consent was requested is (usually) part of ex-ante activities and (usually) is common to a large number of consent requests - such as online consent requests shown to all users on a website. Therefore, assessment of whether it fulfils obligations associated with informed consent is also common to all instances of consent based on it. By performing an evaluation of this artefact in ex-ante phase, its (successful) results can be reused for evaluation of all consent instances based on it in ex-post phase. This represents utilisation of ex-ante test results in ex-post validation of a constraint.

The abstraction of this can be summarised based on considering ex-ante information as that associated with the model or plan of activities, and ex-post information to be regarding execution of those activities. Since the model is a common template for all executions, some common ex-ante validations can be performed prior to execution and results persisted for use in ex-post validations. This information, which are expressed as SHACL validation reports in this particular scenario, are persisted in compliance graph as ex-ante phase validations, and are used as part of data graph in ex-post validations.

An example of an ex-post validation incorporating ex-ante test results is presented in Listing 14. The constraint automatically evaluates outcome of a previous SHACL validation concerning given consent model in ex-ante phase indicated by `sh:ValidationReport` with property `sh:conforms` used by SHACL to indicate whether an evaluated data graph has passed or failed given constraints.

```
:ConsentModelConstraints a sh:NodeShape ;
    sh:targetClass sh:ValidationReport ;
    sh:property :ValidationReportConforms ;
    rdfs:label "Given Consent following Consent Model constraints" .

:ValidationReportConforms a sh:PropertyShape, :AutomaticallyCheckedConstraint ;
    sh:path sh:conforms ;
    sh:hasValue true ;
    sh:message "Consent Model should be compliant for valid given consent" ;
    sh:name "Check validation report says data conforms" .
```

Listing 14: Utilising ex-ante test results for consent model in evaluating ex-post instances of given consent

6.2.4 Proof-of-concept implementation

For a proof-of-concept implementation of the approach, the consent dialogue on the Quantcast[13] website was utilised as a use-case and evaluated for GDPR compliance. Information from consent dialogue was manually analysed and represented using GDPRov and GConsent to create the data graph. Additionally, information from other pages on website was also analysed to identify more information about purposes, processing, personal data, and third parties mentioned in the consent dialogue. Resources associated with the implementation are available in a public repository[14].

The choice of use-case was made based on Quantcast being a provider of GDPR consent collection mechanism using the IAB consent framework[15] - which is the largest consent framework in use and is based on collection of consent and sharing of personal information using the internet. The Quantcast website was also one of the few (at the time and to the authors' knowledge) websites that allowed changing/withdrawing consent using the same dialogue as that used to request/provide it.

The aim of this exercise is to demonstrate use of SHACL in validating information for compliance, and use of ex-ante test results in validating information in ex-post phase. It is not intended to act as a compliance evaluation[16] of Quantcast, but to serve as a demonstration of semantic web in representing, querying, and documenting information for compliance.

6.2.4.1 Description of Consent Dialogue

The consent dialogue, depicted in Figure 6.4, is presented to the user upon visit to Quantcast website. The consent dialogue consists of multiple pages or panels presenting various abstractions of information and choices which the user can interact with. The first panel, depicted in figure as `(a)`, presents a brief description of processing and purposes and pro-

[13] Web archive snapshot https://web.archive.org/web/20190430014325/https://www.quantcast.com/

[14] https://github.com/coolharsh55/GDPR-semantics-demo/

[15] https://advertisingconsent.eu/

[16] The Data Protection Commission of Ireland opened an enquiry on 02-May-2019 into the practices of Quantcast in relation to "processing and aggregating of personal data for the purposes of profiling and utilising the profiles generated for targeted advertising is in compliance with the relevant provisions of the GDPR" - source: https://www.dataprotection.ie/en/news-media/press-releases/data-protection-commission-opens-statutory-inquiry-quantcast. The enquiry was announced well after the completion of the presented work, but bears relevance in terms of its findings - which have not been announced as of February 2020.

vides an option to provide consent[17] using the 'I Accept' button. Further information is made available through the 'Show Purposes' button. Upon giving consent at any stage of dialogue, clicking 'Change Consent' link in footer at bottom of page shows the consent dialogue with previously selected consent choices.

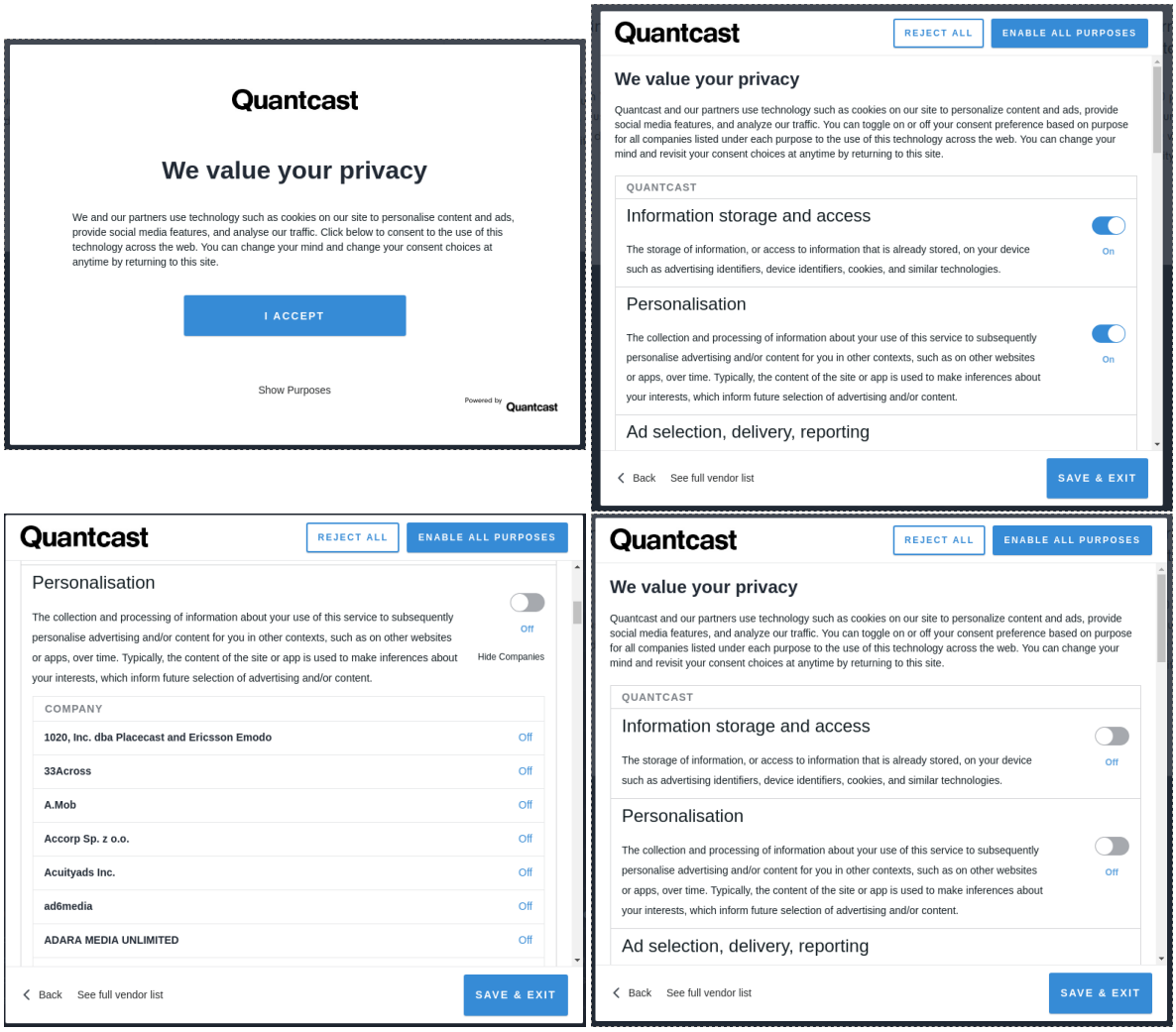


Figure 6.4: Consent dialogues on quantcast.com (clockwise from top-left) (a) first screen (b) default options on selecting "I Accept" (c) default options on selecting "Show Purposes" (d) Third parties listed for purpose "Personalisation"

6.2.4.2 *Extracting Purposes, Processing, and Personal Data from Consent Dialogue*

Clicking 'Show Purposes' dialogue opens a second panel containing information about purposes and third parties associated with consent, displayed in figure as `(b-d)`. The first section provides information about processing of personal data carried out by Quantcast. Its structuring of information consists of title specifying the purpose of processing, for ex-

[17] Note for clarification: Clicking the 'I Accept' button signals consent for all specified purposes, as can be verified by clicking on the 'Change Consent' button at the bottom of the page to show the selected choices in the consent dialogue. We avoided the interpretation of qualitative assessments in evaluating whether the "I Accept" button fails consent requirements such as not having options pre-ticked or pre-chosen by default, though we believe this does not satisfy the requirements of valid consent under GDPR. We instead represent these qualitative constraints as `ManuallyCheckedConstraint` and assume their assessment to always be true.

ample - "*Information storage*", followed by a textual description of personal data categories involved and processing operations to be performed on them.

The purpose was represented as instances of `gdprov:Process` and `gc:Purpose` with the title from consent dialogue specified as its label using `rdfs:label`. Information about processing and personal data categories was manually extracted from text and represented as - `gdprov:Step` and `gc:Processing` for processing, and `gdprov:PersonalData` and `gc:PersonalData` for personal data.

The consent dialogue was represented as `gdprov:ConsentAgreementTemplate` to indicate an ex-ante artefact provided when requesting consent. Since the consent dialogue offers granular choices from which the user can choose any option individually, the question of its semantic representation led to two possible solutions - first where entire consent dialogue and all consent choices are considered a single instance of consent, and second where each individual and independent choice is considered an instance of consent. Since given consent for an individual option in the dialogue could be revoked without affecting other choices, each independent option was chosen to be modelled as an instance of consent. As the consent dialogue acts as a common template for all options, its representation as a 'bundle' of consent templates was added to an updated version of GDPRov (v0.7) as `gdprov:ConsentAgreementTemplateBundle`. This enabled representing a common artefact used to request separate instances of consent. Listing 15 provides an example representation of information from consent dialogue using this data model.

```
:ConsentRequestDialog a gdprov:ConsentAgreementTemplateBundle ;
    rdfs:label "Consent Dialog shown to the user" ;
    rdfs:comment "Dialog that shows - We value your privacy...
        ... customise their choice by clicking on 'Show Purposes'." ;
    gdprov:usesConsentAgreementTemplate
        :CATQInfoStorageAccess, :CATQPersonalise, :CATQAds,
        :CATQMeasure, :CATTPInfoStorageAccess, :CATTPPersonalise,
        :CATTPAds, :CATTPContentSelection, :CATTPMeasure, :CATTPGoogle .

:CATQInfoStorageAccess a gdprov:ConsentAgreementTemplate, gc:Consent ;
    rdfs:label "consent for CATQInfoStorageAccess" ;
    gc:forPurpose :InformationStorageAccess ;
    gc:forProcessing :StoreIdentifiers, :UseIdentifiers ;
    gc:forPersonalData :Cookie, :AdIdentifier, :DeviceIdentifier ;
    gc:hasLocation <https://quantcast.com/> ;
    gc:withdrawBy <https://www.quantcast.com/#displayConsentUI> ;
    gc:inMedium "dialog box on website" ;
    gc:hasStatus gc:ConsentStatusRequested .
```

Listing 15: Representation of consent dialogue as a bundle of consent requests

6.2.4.3 Extracting Third-Parties from Consent Dialogue

In the bottom half of second panel, the consent dialogue provides information about purposes of sharing data with third parties and a list of recipients for each purpose. This can be seen in the figure in panel `(c)`. Consent for each purpose for sharing data with third parties can be individually acted upon by means of a radio button or toggle. Opting to provide consent for a purpose is taken as providing consent for all listed third parties for that purpose, i.e. there is no granular control for consent for individual third parties.

The third parties were defined using `gdprov:ThirdParty` and `gc:ThirdParty`. Although names of purposes are same in sections describing processing by Quantcast (top-half) and by third parties (bottom-half), these were declared as separate instances to reflect separation of choices to provide consent. Third parties were associated with a process by first creating representing the data sharing using `gdprov:DataSharingStep` and `gc:DataStep`, and then using the property `gc:SharesDataWithThirdParty` to link these with the third party.

6.2.4.4 Gathering additional information from Quantcast website

The consent dialogue does not provide information on how personal data categories are collected, or data sources of personal data. To investigate this, an analysis of information about products and services provided by Quantcast on their website along with their policies was carried out to identify relevant information which could be added to complete the use-case.

Measure is a free service offered by Quantcast that provides analytics regarding audience (visitors) to websites. It uses following categories of personal data: *Demographics* (age, gender, family, location, income, education, and occupation), *Psychographics* (purchase history, brand preference, cars driven, media consumption), *Engagement* (categorise visitors as passers-by, regulars, and fanatics), and *Traffic* (platform - web and mobile web, country, time period). Of these, data categories of Demographics and Psychographics were included in data graph as being relevant to information provided in consent dialogue. Their source was not indicated by Quantcast and therefore was not added to data graph. For Psychographics, Quantcast specifies that it uses information from third-parties (Experian, Mastercard, DLX, TiVo, and Netwise) to 'augment' its profiles. The third parties were defined as source for this data based on this information. The profiles mentioned are described on the webpage as broad categories of data in the form of Shopping Interests, Media Interests, Business & Occupation, Geography, and Political Interests. These were added to data graph as personal data categories.

Targeting is a service which provides selecting audiences/users based on personal data attributes (similar to those in Measure). While Quantcast[18] does not explicitly say that it uses the same personal data collected and used in Measure, this was implicit in its description. However, since this is an assumption, it was not included in the data graph.

Measurement is a service similar to Measure and Targeting in its use of audience profile, with the key difference between provision of service beyond website audiences, such as for campaigns. It describes data categories such as Website Traffic, Demographics, Interests, Search behaviours, and Media consumption, which were added to data graph.

The privacy policy provided by Quantcast provides information regarding personal data categories as Cookies, Tags, and Log data - which were added to data graph. The use and collection of emails used to contact Quantcast were also incorporated. Data retention periods are described as "for as long as necessary", with an explicit limit for log data provided

[18]The service provided by Quantcast is in essence similar to that provided by Facebook - it acts as the mediator between providers and consumers by matching the criteria to user profiles. For example, it mentions an example where the target audience is "women 18-34 who love shopping, travelling and wine", which implies that it must know about a) gender b) age range c) website history d) purchase history e) travel history, It further elaborates, "We build a custom model based on millions of available data points about your audience, such as their pre-search behaviours, demographics, and past purchases."

as 13 months. Due to the ambiguity and pending legal resolution of temporal limits, this information was not added to data graph. The privacy policy also described GDPR rights regarding right to access, right to rectify, right to restrict processing, right to deletion, right to data portability, and provided a link[19] for contact and more information. This link was used as the IRI for activities associated with these rights.

6.2.4.5 Validating using SHACL

As the use-case concerns consent, constraints associated with consent in Section 4.2.3 were used to validate information using the approach described in Section 6.2.1. For evaluation, three sets of constraints were developed to validate: (a) only ex-ante model of consent dialogue, (b) instances of given consent, and (c) reusing results of ex-ante consent dialogue tests to validate given consent. This allowed an analysis and comparison of combining ex-ante and ex-post validations, and to demonstrate benefits in terms of reduced validations and reuse of compliance information.

SHACL constraints were executed using the TopBraid SHACL binary[20]. A bash[21] script enabled automation in execution of constraints as different approaches (ex-ante, ex-post, combination of both) and persistence of test results as separate files.

For ease of evaluation, a combined data graph was created consisting of data from Quantcast and ontologies used - GDPRov, GConsent, GDPRtEXT. The data graph and test results were enhanced (and verified for logical consistency) using a semantic reasoner[22] to identify and add additional triples derived from inferences. The resulting data was added to a triple store[23] in separate graphs representing data graph and compliance graph.

6.2.5 Generating reports using SPARQL

The triple store enabled querying of information to generate compliance reports and documentation. The use of GraphDB provided access to some in-built reasoning capabilities[24] which were useful in the querying process. While a number of SPARQL queries were constructed based on compliance questions and are available for introspection in the code repository, only one is provided here as an example to demonstrate retrieval of information and documentation of compliance information.

The SPARQL query, listed in Listing 16, retrieves information about each tested validation constraint, its result, link to GDPR, and whether it passed or failed. The results, shown in Table 6.3 act as a test report, and contain constraint description (Name), type - automatic (A) or manual (M), link to GDPR, result - pass (P) or fail (F), and node (instance in data graph) if it failed a constraint. The report also contains a failure message associated with the constraint that is not shown in table due to space limitations.

The rows which correspond to failed constraints are manually highlighted to provide an indication of information in a visual medium - such as a dashboard. The query and its results

[19] NOTE: The rights information page could not be accessed in this case with the webpage providing an error regarding Quantcast cookies not being set."
[20] https://github.com/TopQuadrant/shacl
[21] https://www.gnu.org/software/bash/
[22] HermiT http://www.hermit-reasoner.com/
[23] GraphDB Free Edition http://graphdb.ontotext.com/
[24] http://graphdb.ontotext.com/free/devhub/inference.html

```
PREFIX c: <http://example.com/Quantcast/shapes#>
PREFIX sh: <http://www.w3.org/ns/shacl#>
SELECT DISTINCT ?name ?test ?gdpr ?result ?node ?msg
WHERE {
    ?x a c:Constraint .
    ?x sh:name ?name .
    BIND(
        IF(EXISTS{?x a c:AutomaticallyCheckedConstraint},
            "Automatic"^^xsd:string, "Manual"^^xsd:string)
        as ?test)
    OPTIONAL { ?x c:linkToGDPR ?gdpr }
    BIND(
        IF(EXISTS{?y sh:sourceConstraint ?x},
            "FAIL"^^xsd:string, "PASS"^^xsd:string)
        as ?result)
    OPTIONAL {
        FILTER EXISTS { ?y sh:sourceConstraint ?x } .
        ?y sh:focusNode ?node .
            ?y sh:resultMessage ?msg . }
} ORDER BY ?name
```

Listing 16: SPARQL query for report listing validation results linked with GDPR

can both be persisted in machine-readable serialisations using standards for representations, making them interoperable and capable of automation. The information derived from such validations and querying is useful to generate compliance documentation and reports for an organisation to oversee their compliance with GDPR - which is itself an obligation mandated by GDPR.

Table 6.3: SHACL validation report linked to GDPR

Name	Type	GDPR	Result	Node
Consent ≠ Inactivity	M	R32	P	
Consent ≠ Pre-ticked Boxes	M	R32	P	
Consent ≠ Silence	M	R32	P	
Consent → Data Subject	A	A4-11	P	
Consent → Given To	A		P	
Consent → Location	A		P	
Consent → Medium	A	A7-2	P	
Consent → Personal Data	A	A4-11,R32	P	
Consent → Processing	A	A4-11,R32	P	
Consent → Provided By	A	A7-2	P	
Consent → Purpose	A	R32,R42	P	
Consent → Status	A		P	
Consent → Timestamp	A		F	Q:Consent20190415120753
Consent → Timestamp	A		F	Q:Consent20190415140000
Consent ≡ Choice	M		P	
Consent ≡ Freely Given	M	A4-11	P	
Consent ≡ Specific	M	A4-11	P	
Consent ≡ Statement of Clear Action	M	A4-11	P	
Consent ≡ Unambigious	M	A4-11	P	
Consent Generating Activity	A		P	

(Cont'd on following page)

SHACL validation report linked to GDPR (cont'd)

Name	Type	GDPR	Result	Node
Consent Request ≡ Clear	M	R32	P	
Consent Request ≡ Concise	M	R32	P	
Consent Request ≡ Not Disruptive	M	R32	P	
Consent Template	A		P	
Ease of Withdraw Consent	M	A7-3	P	
Many Processing x One Purpose	A	R32	P	
One Processing x Many Purposes	A	R32	F	Q:Consent20190415120753
One Processing x Many Purposes	A	R32	F	Q:Consent20190415140000
Personal Data → Storage Period	A	A13-2-a	F	Q:CATQInfoStorageAccess
Personal Data → Storage Period	A	A13-2-a	F	Q:CATTPInfoStorageAccess
Personal Data → Storage Period	A	A13-2-a,R39	F	Q:Consent20190415120753
Personal Data → Storage Period	A	A13-2-a,R39	F	Q:Consent20190415140000
Right to Withdraw	A	A7-3	P	
Separation of Processing	M	R43	P	
Third Party Categories	A	A44	P	
Third Party Identities	A	A13-1-e	P	
Third Party Identities	A	A30-1-d	P	
Third Party Identities	A	A44	P	
Third Party Safeguards	A		P	
Withdraw Consent Information	M	A7-3	P	

6.2.6 Evaluation

The approach described in Section 6.2.1 for the conceptual model has been published as a peer-reviewed publication [59] in Poster & Demo track at the SEMANTiCS conference in 2018 - which involves a good mix of industry and academic participants and thereby provided opportunity to present this work to industry community. The construction of a knowledge graph based on evaluations of GDPR compliance was published as a peer-reviewed publication [58] in workshop on Contextualized Knowledge Graphs which was co-located with International Semantic Web Conference (ISWC). The workshop provided reviews and feedback from domain experts regarding use of semantic web to create knowledge graphs, and how it could be utilised in legal compliance domain. The proof-of-concept implementation presented in Section 6.2.4 was published as a peer-reviewed publication [60] at the SEMANTiCS conference in 2019.

Effectiveness of combining ex-ante and ex-post validations

In order to understand number of validations in testing process, consider the set V_t consisting of all validations that should be evaluated in order to determine validity of given consent. This set consists of validations evaluating information in consent dialogue which is common to all instances of given consent - represented by V_a. The remaining validations consist of evaluating information specific to an instance of given consent, such as timestamps, and are represented by V_p. To summarise, set of validations consists of validations

evaluating the consent dialogue and information associated with given consent, giving the expression $V_t = V_a + V_p$.

V_a is required to be carried out as part of ex-ante compliance evaluations where the organisation must monitor and ensure its activities are compliant before any processing takes place. In this case, the consent dialogue box is required to be evaluated and found compliant before any consent is requested. Therefore, V_a represents ex-ante validations and V_p represents ex-post validations.

If results of V_a are persisted, then they can be reused in evaluation of given consent by simply checking whether outcome of V_a was valid or invalid - in a single validation. Therefore, total validations to be performed when combining ex-ante and ex-post validations is $V_t = 1(V_a) + V_p$ - which is efficient assuming $V_a > 1$.

In the use-case of consent dialogue presented in this section, $V_t = 59$ validations of which $V_a = 57$ validations and $V_p = 2$ validations. If all validations were evaluated for given consent, each instance of given consent would need 59 validations. Whereas, if the ex-ante validations were reused and only the ex-post validations were evaluated, then each instance of given consent would need only 3 validations to be evaluated (1 validation to evaluate $V_a + 2$ validations from V_p). While these numbers are use case specific, it clearly demonstrates that the approach is more efficient in terms of validations and determining validity of given consent. This is assuming the ex-ante model of consent dialogue was found compliant in ex-ante stage, and therefore its validation only evaluated presence of a test report affirming its compliant status.

Comparison with SotA

Table 6.4 provides a comparison of use of SHACL with approaches within SotA based on earlier analysis in Section 3.7.6. The SPECIAL project uses a semantic reasoner to determine whether a given combination of processing operations expressed as OWL2 class axioms are valid [92], while work presented by Vos et. al [62] uses ODRL profiles to express requirements which are converted to and evaluated using Answer Set Programming (ASP). The MIREL project detects violations of GDPR by utilising the PrOnto ontology [41], [55], [63] to model legal concepts and LegalRuleML to model norms, which are then applied over a BPMN use-case using Regorous to generate a report [63]. The DAPRECO project uses PrOnto along with Reified Input/Output logic (RIO) [121] to specify norms and rules to create a knowledge base [118] which is then used to identify relevant obligations to check for compliance. These efforts show the variety in evaluation approaches for compliance and the use of semantic web technologies in evaluation of compliance.

The work described in this section demonstrates how SHACL can be used to validate information for correctness and adherence to obligations mandated by GDPR based on interpretation of compliance questions from Chapter 4. In this form, SHACL can be used to evaluate compliance, though presented work focused on validation of constraints based on concept of compliance questions. Compared to state of the art in Chapter 3, the work regarding SHACL (highlighted first row of table) is novel within SotA in use of SHACL and linking of results to GDPR in a machine-readable and thus query-able form. In addition, creation of a compliance graph to store information associated with demonstration of compliance enables using SPARQL queries to identify remedial measures to achieve compliance

as well as create reports to identify and present information relevant to compliance. The utilisation of ex-ante test results in ex-post validations is also novel within state of the art and provides an efficient method for validation of compliance information.

In comparison with SHACL, approaches in SotA use or advocate formal methods based in logic where legal norms can be expressed in terms of requirements and evaluated to determine compliance. In turn, SHACL provides a validation framework where results can be persisted as a graph and queried. In addition, SHACL validations can be linked to GDPR, as demonstrated using GDPRtEXT, which makes it possible to use SHACL to verify the output of other compliance evaluation approaches and record their outcomes as a test result linked to GDPR clauses, thereby creating reports of compliance. This also provides an opportunity to explore reuse of existing approaches and resources regarding evaluation of GDPR compliance where SHACL is used to generate an interoperable overview of evaluation results while abstracting underlying outputs from different approaches. For this, resources provided by Vos et. al [62] and DAPRECO project [118] provide constraints expressed using logic-based formalisms that are available as open access, providing future direction for applicability of this research.

Table 6.4: Comparison of SHACL validation with SotA

Approach	**Evaluation method**	**Scope**	**Machine-readable result?**	**Provides remedies?**	**Links results to GDPR?**
Pandit	SHACL	RDF data	✓	✓	✓
SPECIAL	OWL	Consent	✓		
SPL+SERAMIS	ODRL	Obligations	✓	✓	✓
SPL+Vos et al.	OWL, ASP	Obligations	✓	✓	
SPL+CitySPIN	OWL	Consent	✓		
MIREL	RuleML	Obligations	✓	✓	✓
MRL+DAPRECO	RuleML	Obligations	✓	✓	✓
BPR4GDPR	OWL	Process Flows		✓	
Lodge et al	SDK	Process Flows	✓	✓	
Tom et al	BPMN	Process Flows	✓	✓	
Corrales et al	Questionnaire	Obligations			
LUCE	Smart Contracts	Data Sharing	✓		
AdvoCATE	Smart Contracts	Consent	✓		
Sion et al	UML, DFD	Process Flows	✓	✓	
privacyTracker	Access Control	Data Sharing	✓		
Robol et al	STS	Process Flows	✓		
GuideMe	Questionnaire	Process Flows		✓	
Basin et al	Algorithm	Process Flows			
RestAssured	XACML	Process Flows	✓		
DEFeND	Questionnaire	Obligations	✓		
OPERANDO	Access Control	Process Flows	✓		
PoSEID-on	Smart Contracts	Data Sharing	✓		
DECODE	Smart Contracts	Consent	✓		

SUMMARY

Summary of work presented

Section 6.1 presented use of SPARQL queries in representing compliance queries and retrieving information associated with compliance that was represented using GDPRov and GDPRtEXT ontologies. The work fulfilled research objective $RO4$ by representing compliance questions as SPARQL queries and demonstrating their application using a real-world use-case. The demonstrated application used questions from GDPR readiness guide published by Data Protection Commission of Ireland in 2017 to assist organisations in assessing their adherence to compliance requirements of GDPR. The created SPARQL queries retrieved information for answering these compliance questions for a synthetic use-case based on the scenario of an online shopping service. The queries and the demo were published in a peer-reviewed publication [57] at SEMANTiCS conference and are available online as an application with resources provided in a code repository.

Section 6.2 presented use of SHACL to validate information regarding its correctness and adherence to obligation towards GDPR compliance. This work fulfilled research objective $RO5$ by validating information using SHACL and linking results with relevant clauses of GDPR for compliance documentation. The SHACL validation utilised constraints developed from analysis of compliance questions as presented in Section 4.2.3. An approach for the validation process using SHACL was presented in which ex-ante test results were reused in validation of ex-post information. This enabled efficient evaluations by reducing number of validations required in ex-post phase, and also enabled associating compliance of ex-ante information with that of its corresponding ex-post information. A demonstration of the approach was provided through a use-case in which the consent dialogue on a real-world website was represented using developed ontologies and validated using SHACL. The results of validation were queried using SPARQL to generate documentation for compliance in the form of a test report which showed compliance status of different obligations and highlighted failing tests as action items for meeting compliance requirements. The approach of using SHACL and combination of ex-ante and ex-post validations was published in a peer-reviewed publication [58] at the SEMANTiCS conference in 2018, while the demonstration was published in a peer-reviewed publication [60] at SEMANTiCS 2019. The resources associated with the work have been made available online in a code repository.

This chapter, through both presented works, provides an application of developed ontologies presented in Chapter 5 for querying and validating information about GDPR compliance. It serves to demonstrate usefulness of these ontologies, and provides an indication of their role in representation of information. The chapter also demonstrates use of semantic web technologies in representing, querying, and validating information for GDPR compliance.

The modular test-based approach can be used with existing representations in non-RDF data that are evaluated using other tools and methods by adding semantics to test results and reports to link them with relevant information in GDPR. This will enable utilisation of a validation method such as SHACL to evaluate its correctness and a querying method such as SPARQL to retrieve information in the form of compliance test reports.

The advantages of representing processes with semantics go beyond testing for compli-

ance as representation of processes are also useful for planning of operations and internal documentation. Semantic representations of processes can assist in automating the generation of documentation such as privacy policies where processes are listed along with their purpose, legal basis, and use of personal data. Privacy policy generators that generate boilerplate policies exist online, but do not currently incorporate semantics. The use of semantics allows query-able machine-readable metadata that can be used in tools towards understanding and evaluating policies for users and authorities.

Re-usability of developed resources

The interpretation of compliance questions in GDPR readiness document using SPARQL and its application in synthetic use-case demonstrates the potential application and usefulness of SPARQL queries to retrieve information relevant for compliance. However, it also showcases that creation of SPARQL queries is highly dependant on utilising the same ontological concepts as the data it is querying. Therefore, such SPARQL queries are ontology-dependant, and by definition do not have re-usability beyond the data they were created for. The same is true for constraints represented in SHACL, which are dependent on the underlying ontologies used to represent the data graph it intends to validate.

Using the analysis and natural language basis of the questions and constraints, another approach can adopt these resources to query and validate information using its use-case specific ontologies. While the individual query or constraint would need rewriting, the overall approach and modelling of tests can be reused to generate similar compliance documentation. The provision of all resources under permissive licenses provides an adopter with access to underlying data and information to assist them in this process.

Novelty of presented work

The use of SPARQL and SHACL for GDPR compliance as presented in this chapter is novel within state of the art as presented in Section 3.7. While approaches in SotA use SPARQL to query information, none present their work as intended to answer compliance questions or as intended to investigate compliance of an organisation. The use of SHACL is a first within SotA regarding validation of information for GDPR compliance based on analysis of approaches in Section 3.7.6. In addition, work presented in this thesis has been published in peer-reviewed publications with open access to its data and resources for transparency. Together, these serve as novel contributions that extend the state of the art.

7 | CONCLUSION

This chapter concludes the thesis with a discussion on the extent to which the research question (Section 1.2) and objectives (Section 1.2) have been addressed through presented work. The chapter also presents resulting contributions which were previously summarised in Section 1.4, and its comparison with similar approaches in state of the art (SotA) identified in Chapter 3. The chapter concludes with potential avenues for further work arising from research presented within the thesis.

7.1 FULFILMENT OF RESEARCH OBJECTIVES

The research question guiding the work presented in this thesis, defined in Section 1.2, is - *"To what extent can information regarding activities associated with processing of personal data and consent be represented, queried, and validated using Semantic Web technologies for GDPR compliance?"*. Five research objectives were identified which guided the work towards answering the research question. This section discusses the extent of their fulfilment based on work presented in previous chapters of the thesis.

Fulfilment of RO1 and RO2

The first two research objectives ($RO1$ and $RO2$) were concerned with identifying the subset within GDPR regarding activities associated with personal data and consent, and the information required for evaluating its compliance. This was fulfilled by work presented in Chapter 4.

The first research objective ($RO1$) required identification of a subset of GDPR relevant to activities associated with personal data and consent in ex-ante and ex-post phases of compliance. The second research objective ($RO2$) was to identify information required to represent activities associated with personal data and consent for GDPR compliance based on the identified clauses of the GDPR from $RO1$. To facilitate this, an information model was developed to explore the entities and their relationships with respect to information exchange guided by GDPR compliance requirements (Section 4.1). The model provided an analysis of GDPR in the form of requirements and processes associated with information for compliance, and was used to categorise the requirements of information as Provenance, Agreements, Consent, Certification, and Compliance. These categories were then used to analyse and identify the nature and source of information required for compliance, and its relationship with entities and stakeholders defined by the GDPR.

The information requirements were expressed in the form of questions, termed 'compliance questions' (Section 4.2), which provided structure for identifying information required

to answer them for evaluating compliance. Authoritative sources used in the information gathering process for $RO1$ and $RO2$ included European Data Protection Commissioner's offices, reports and opinions produced by Article 29 Working Party regarding interpretation of GDPR, information about case law pertaining to the interpretation of the GDPR, and documents published by institutions providing legal services.

As the motivation for work was utilisation of semantic web to represent this information, the methodology of using 'competency questions' was adopted to enable the formulation of an ontology from identified information [35]. This was done by interpreting the compliance questions as 'competency questions' and identifying concepts and relationships about activities associated with personal data and consent to answer the questions (Section 4.2). From these compliance questions, a set of constraints were identified which the information needed to satisfy in order be valid for its use in evaluating compliance, and the assumptions which always hold true (Section 4.2.3). These constraints and assumptions were utilised in Chapter 6 for development of an approach for validation of information as required to fulfil research objectives $RO4$ and $RO5$.

RO3

The third research objective ($RO3$) was to create ontologies to represent information identified in $RO2$ about activities associated with personal data and consent in ex-ante and ex-post phases for GDPR compliance. $RO3$ was divided into three sub-objectives, which involved creation of ontologies for representing - (a) concepts and text of GDPR, (b) activities associated with personal data and consent, and (c) information regarding consent. The objective was fulfilled through work presented in Chapter 5 consisting of GDPRtEXT, GDPRov, and GConsent ontologies for each of respective sub-objectives.

The ontologies were developed following well established methodologies [34]–[36] and best practices advocated by the semantic web community, as summarised in Section 1.3.3. They were evaluated based on their ability to represent information required to answer the competency questions [35] from $RO2$, as well as against common pitfalls in design using the OOPS! tool [38]. The ontologies were documented using the WIDOCO tool [37], utilised persistent identifiers provided by W3ID in name-spaces, and were published under an open license in the public repository at Zenodo with DOIs. Each ontology was compared against the state of the art to identify the extent of information representation and novelty of its concepts and approach.

The first sub-objective ($RO3(a)$) was fulfilled with the GDPRtEXT ontology (Section 5.2), which enabled unambiguous and machine-readable linking of information to concepts and text of GDPR. GDPRtEXT provided an OWL2 ontology to represent the structured text of GDPR as individual Recitals, Chapters, Sections, Articles, Points, Sub-Points, and Citations, by extending the European Legislation Identifier (ELI) ontology [42]. ELI is the authoritative ontology used by the European Publication Office to define metadata for all published documents. The extension mechanism used by GDPRtEXT maintains formal compatibility with ELI. Using GDPRtEXT, the text of GDPR was re-defined as linked data in machine-readable representations by assigning an unique identifier for individual resources, which made it possible to define machine-readable links to specific clauses of the GDPR. In addition, a thesauri of terms and concepts defined or referenced by the GDPR was provided using SKOS.

GDPRtEXT thus fulfils $RO3(a)$ regarding provision of a mechanism for associating information with concepts and text of the GDPR.

The second sub-objective ($RO3(b)$) was fulfilled by the GDPRov ontology (Section 5.3) by enabling representation of activities associated with personal data and consent in ex-ante and ex-post phases. GDPRov extends the PROV-O [47] and P-Plan [48] ontologies with terms and relationships relevant for GDPR, where PROV-O is the W3C standard for representing provenance information, and P-Plan is its extension for defining abstract models as plans which then get instantiated into activities having provenance. GDPRov extends the PROV-O and P-Plan ontologies to represent a model or plan of how processes are supposed to interact with personal data and consent (ex-ante phase), such as for collection, use, storage, and sharing. The model or plan can then be used as the template for activities to be carried out whose provenance (ex-post phase) is linked to the model. Apart from providing terms for addressing personal data and consent, GDPRov also enables representation of other activities defined by GDPR, such as handling rights and data breaches, which can similarly be depicted using a model or plan.

The third research sub-objective ($RO3(c)$) was fulfilled by the GConsent ontology (Section 5.4) by enabling representation of information associated with consent. GConsent expands upon the use of consent as an abstract entity in GDPRov by providing representation of contextual information associated with actors, state, relationships, and provenance of consent as required for compliance. In particular, it provides representations for association of purpose, processing, personal data, data subject, third parties, and delegates with a specific instance of consent. It also provides representations for contextual information such as the medium the consent was given, timestamp, and location. GConsent also provides the novel notion of 'states' which reflect the status of consent for compliance and provide an indication of its use, such as 'requested' or 'explicitly given' or 'invalidated', which are categorised based on whether they can be used as a valid legal basis for processing. GConsent thus provides a comprehensive ontology for the representation of information associated with consent for GDPR compliance and fulfils $RO3(c)$.

RO4 and RO5

The fourth research objective ($RO4$) was to create SPARQL queries that retrieve information about activities associated with personal data and consent for GDPR compliance. These SPARQL queries were formulated as semantic representations of compliance queries identified in $RO2$ and utilised ontologies created in $RO3$ to define concepts and relationships pertaining to GDPR. The SPARQL queries demonstrate the linking of retrieved information with relevant concepts and parts of the GDPR, as well as the creation of knowledge graphs for use in compliance processes. This work was presented in Chapter 6.

The fifth and final research objective ($RO5$) was the creation of a framework utilising SHACL to validate information regarding activities associated with personal data and consent, and linking the results to relevant concepts and clauses of GDPR. This was fulfilled by the work presented in Chapter 6. The framework utilised SHACL to validate information based on the constraints and assumptions identified in $RO2$ and presented in Chapter 4. The validation tests utilised the developed ontologies in $RO3$ to define the concepts and relationships, and to annotate the test and their results with links to relevant clauses of the

GDPR. The framework also utilised SPARQL queries generated in $RO4$ to retrieve and validate information by using SHACL-SPARQL.

The framework was demonstrated and evaluated through a use-case generated from the consent mechanism on a real-world website, where the information associated with consent was validated for ex-ante and ex-post phases of compliance. Information about the consent mechanism was represented using the developed ontologies (GDPRov and GConsent), with SHACL used to represent ex-ante and ex-post validation tests. The ex-post approach validated individual instances of consent from provenance log of given consent, while the ex-ante approach validated the template used to provide information and choices for requesting consent. A third approach was developed which utilised a combination of both ex-ante and ex-post approaches by validating common requirements on the consent template in the ex-ante, and persisting its results for reuse in the ex-post validation of unique validations for given consent in the provenance log. The combined approach was shown to be more efficient in terms of reducing the number of validations as compared to individually validating ex-ante and ex-post requirements.

The SHACL tests defined for validation were annotated with an additional property that linked them with the relevant clauses and concepts of the GDPR using GDPRtEXT. This property was used to associate the validation test and result with the GDPR, and provided the basis for querying information regarding validation against GDPR clauses and concepts. The framework thus demonstrated the validation of information for GDPR compliance regarding activities associated with personal data and consent, and linking the validation results with relevant clauses and concepts of the GDPR using GDPRtEXT, thus fulfilling $RO5$.

7.2 EXTENT OF SEMANTIC WEB TECHNOLOGIES IN ADDRESSING RQ

The research question guiding this thesis focuses on the representation, querying, and validation of information as the basis upon which GDPR compliance is evaluated. More specifically, it concerns activities associated with processing of personal data and consent - which, while being an important part of GDPR, represents only a subset of requirements in the GDPR. To investigate the 'extent' aspect in the research question, the research objectives were formulated to correspond with information representation ($RO3$), querying ($RO4$), and validation ($RO5$). In addition to these, the use of linked data principles enables associating information - in general and including queries, validations, and results - with clauses of the GDPR. This provides the argument for specifying that semantic web technologies provide information management with respect to its representation, querying, validation, and association with the GDPR in the process of compliance.

While the above is sufficient to cover the scope of the thesis, the domain of GDPR compliance (and legal compliance in general) has other areas where semantic web has been demonstrated to be capable from analysis of the state of the art. Existing approaches have demonstrated use of semantic web technologies in representing information regarding compliance as deontic logic, norms, requirements, and other logic based formalism which are used to represent the requirements of GDPR and are evaluated as a measure of compliance itself (see SotA in Chapter 3). These serve to prove that there is more than one way to represent and evaluate information towards evaluating GDPR compliance using semantic web.

At the same time, there is a lack of information representation and approaches utilising semantic web technologies for addressing the larger scope of information associated with GDPR compliance as discussed in Section 4.1 regarding data governance, data processing agreements, and documentation of information. The existing approaches - including the contributions of this thesis - provide the necessary building blocks for addressing representation of information required to evaluate compliance, expressing information about compliance itself, querying information, validating it for correctness, evaluating information for sufficiency to an expressed obligation or requirement, compiling reports or records, recording provenance of activities, and generating documentation for compliance. Semantic web is notable in providing a consistent, coherent, interoperable, and modular set of technologies for carrying out the above activities - which makes their use in development of legal compliance solutions particularly attractive due to the complexity of the domain and a need to expand or specialise applications in use-cases. Therefore, from the perspective of information and knowledge modeling - semantic web provides the foundational set of technologies useful towards carrying out the activities associated with GDPR compliance.

Future work within this domain largely consists of utilising existing approaches towards extending or revising existing work in an application-oriented manner with a few examples of these mentioned as future work in Section 7.4.

To end the discussion with an analogy - *Though all roads lead to Rome, we aren't there yet.*

7.3 Contributions

This section provides a summary of contributions arising from the research presented in this thesis, which were initially summarised in Section 1.4. The thesis yielded two major contributions - using semantic web to enable linking of information with concepts and text of GDPR, and ontologies for representing information about activities associated with personal data and consent for GDPR compliance. The thesis also yielded minor contributions in the form of an information model for interoperability between entities associated with the GDPR, and a framework for querying and validating information for compliance using semantic web technologies. The impact and extent of the contributions in terms of publications was listed in Section 1.4.7, which included 17 publications related to the work presented in this thesis. The impact and relevance of the work presented in this thesis also includes participation in the W3C Data Privacy Vocabularies and Controls Community Group (DPVCG) and its deliverable - the Data Privacy Vocabulary (DPV), as elaborated in Section 1.4.6.

Major Contributions

GDPR as a Linked Data Resource

The first major contribution, represented by GDPRtEXT, enables association of information with the concepts and text of GDPR using linked data principles. It provides machine-readable unique identifiers for each specific part (Chapter, Article, clauses etc.) of the GDPR by representing its text in RDF using an extension of the European Legislation Identifier (ELI) ontology. It also provides a SKOS vocabulary of concepts and terms defined or represented within the text of the GDPR. The usefulness of GDPRtEXT has been demonstrated

in its use to define the source of terms in the ontologies presented in this thesis, as well as in linking information related to compliance with the relevant concepts and clauses of the GDPR.

GDPRtEXT advances the state of the art in its provision of unambiguous and machine-readable representations of concepts and text of GDPR (see Section 5.2.4). It is currently the only ontology addressing GDPR that extends ELI, and the only open ontology for concepts associated with the GDPR [43]. GDPRtEXT is also the only approach providing a glossary of terms associated with GDPR compliance. Its use and extension of ELI has had an impact on the development plans of the ontology by the EU Publications Offices by demonstrating the use of granular representation of legal clauses and the necessity of linking terms with their occurrences and definitions within the text. It also had an impact on the creation of the DPV by providing a vocabulary of concepts linked to their definition and use in the GDPR.

GDPRtEXT has received 19 citations to date (excluding self-citations), and has been referenced by approaches in the SotA in context of modelling GDPR concepts. GDPRtEXT is available under an open license (CC-by-4.0) along with its documentation at https://w3id.org/GDPRtEXT/, and has been incorporated into Ireland's open data portal as an dataset with 5 star rating for satisfying linked data principles.

Ontologies for representing activities associated with personal data and consent

The second major contribution is the GDPRov and GConsent ontologies, which together enable representation of information about activities associated with personal data and consent relevant for investigation of GDPR compliance.

GDPRov extends the existing ontologies of PROV-O and P-Plan with concepts and relationships specific to GDPR in order to represent provenance of personal data and consent at ex-ante and ex-post stages. Where ex-post representations are common as provenance logs, the ex-ante representations act as a model or plan or template of intended activities for evaluation of compliance. Furthermore, provenance logs (ex-post) can be linked to their models (ex-ante) to represent the relationship between planning and implementation of processes within an organisation. GDPRov also enables representation of other activities associated with the GDPR such as the handling of rights and data breaches.

Compared to the SotA (see Section 5.3.4), GDPRov provides the most comprehensive representation of concepts and relationships for activities associated with GDPR. It is also the only ontology to provide ex-ante and ex-post concepts within the same representation. To date, the publication of GDPRov has received 18 citations (excluding self-citations). It has been used in an approach to model data flow diagrams (DFDs) for analyses of compliance [183]. GDPRov has been released under an open license (CC-by-4.0) and is available along with its documentation at https://w3id.org/GDPRov/.

GConsent expands upon the abstract representation of consent in GDPRov to provide more verbose information regarding entities and contextual information relevant for the management of consent. It also provides the concept of 'consent states' which reflect the use of consent as a valid legal basis and are useful in the representation and management of consent in information systems. To date, GConsent is the most comprehensive vocabulary regarding consent associated with the GDPR (see Section 5.4.5). GConsent had a direct impact on the representation of consent in the DPV by providing the concepts and compe-

tency questions associated with consent based on GDPR requirements. GConsent has been released under an open license (CC-by-4.0) and is available along with its documentation at https://w3id.org/GConsent/.

Together, the three ontologies (GDPRtEXT, GDPRov, and GConsent) enable the representation of activities associated with personal data and consent for GDPR compliance, and to link information represented using them with the clauses of the GDPR. This enables the use of metadata to annotate legal documents, and automation in the management of information by utilising aspects of querying and validation in the governance process.

Minor Contributions

The minor contributions of this thesis are - an information model of entities and their relationships defined by the GDPR, and a framework utilising semantic web technologies for validating and evaluating information for GDPR compliance. The minor contributions complement the previously described major contributions by providing a theoretical basis in the form of an information model, and demonstrate the feasibility and usability of developed ontologies through an application for validating information for compliance.

The first minor contribution is an information model, which was presented in Section 4.1, provides an analysis of information exchanged between entities and its interoperability based on requirements of GDPR. It provides a categorisation of the information requirements as provenance, agreements, consent, certification, and compliance, and the exploration of existing standards in representation these in an interoperable form for GDPR compliance. The information model advances the state of the art by being the first systemic analysis of information flows and interoperability associated with the entities and stakeholders within the context of GDPR compliance. The model serves to identify and evaluate the potential applications of technology in addressing requirements, and provides motivation to the argument for using semantic web as a suitable representation based on the notion of semantic interoperability.

The second minor contribution, presented in Chapter 6, is the utilisation of semantic web technologies to query information for GDPR compliance using SPARQL and the developed ontologies - GDPRtEXT, GDPRov, and GConsent - to represent the compliance questions as queries that are executed over data represented using the developed ontologies. The approach provides assistance with the investigation of compliance by providing an automated way to query required information. This was demonstrated through the use of SPARQL to represent questions from templates provided by Ireland's Data Protection Commission for assisting organisations with their GDPR compliance. Compared to the SotA, the approach is novel in its use of authoritative sources for compliance questions, and the linking of information with GDPR using GDPRtEXT.

The third minor contribution, as presented in Chapter 6, is a framework that utilises SHACL to validate information for GDPR compliance and link the results with relevant clauses of the GDPR using GDPRtEXT. The framework enables the creation of machine-readable metadata associated with the GDPR, which in turn makes it possible to automate the generation of documentation regarding assessment of compliance. The demonstration of the approach, conducted on a consent mechanism from a real-world website, demonstrate its use in validating both ex-ante and ex-post phases. In addition, the demonstration

also provides the advantages of combining the ex-ante and ex-post phases to create a more efficient compliance mechanism by abstracting the common validation tests to the ex-ante phase and validating only the unique constraints associated with an instance in the ex-post phase. The framework advances the SotA through its novel use of SHACL for GDPR compliance, the combination of ex-ante and ex-post phases of validation, and the linking of validation results with the clauses of the GDPR to create machine-readable documentation for compliance.

Contributions to the DPVCG

The ontologies presented in this thesis - namely GDPRtEXT, GDPRov, and GConsent - were used as an input by the W3C Data Privacy Vocabularies and Controls Community Group (DPVCG) in its analysis of existing work towards creating a standardised common vocabulary. In addition, the author of the thesis was an active contributing member towards the development of the Data Privacy Vocabulary (DPV), and served as the editor for its specification. The DPV provides an ontology associated with personal data processing and legal compliance, including GDPR, and represents a community consensus regarding its definitions, usage, and representation. It is available and documented at http://w3.org/ns/dpv.

Comparing the DPV with the ontologies presented in this thesis, the DPV provides a high-level abstraction whereas the ontologies in this thesis - GDPRov and GConsent - represent a more comprehensive and detailed model for representation of information, making them complimentary in usage with the DPV.

7.4 Opportunities for Further Work

Due to the novelty of GDPR and increased interest in its compliance, there are several opportunities where the work presented in this thesis can be further developed and applied, as categorised in the following three areas -

Align approaches for Regulatory Compliance

Differences in domain ontologies offer varying perspectives on the modelling of relationships and concepts within the same domain. In the case of GDPR, these ontologies can be compared using the commonality of concepts and aims. For example, 'consent' is represented in the ontologies GDPRtEXT, GDPRov, GConsent, SPECIAL [85], and PrOnto [55] - where each representation is based on the same concept of consent, and yet differs in its modelling of the relationships associated with consent. A comparison of ontologies based on semantics of concepts is useful to establish compatibility in their usage and approaches, and to evaluate their usefulness for a given use-case.

The state of the art, presented in Chapter 3, describes existing work outlining such an analysis [43] and its use in a tool [188] to compare approaches in the legal domain. It also presents approaches involving application of deontic logic to address regulatory compliance for GDPR, where the text is interpreted using ODRL [40], and PrOnto [55] which models deontic operations and uses LKIF [119] to model actions and roles. This can be further expanded to align existing approaches and ontologies through semantics of concepts provided by GDPRtEXT.

Expand Scope of Ontologies

Incorporate future updates to ELI into GDPRtEXT

GDPRtEXT addresses the aim of linking to specific parts of the GDPR by extending the ELI ontology. The EU Publications Office, as the official developers and maintainers of ELI, are currently working on updating the ELI ontology to enable such linking for all published documents. Their work will provide authoritative URIs for all aspects of a legal document, and will also enable identification of definitions. Once published, the updated ELI ontology will make the GDPRtEXT extension redundant. However, GDPRtEXT will still have uses as a SKOS vocabulary of concepts that is used by ontologies such as GDPRov and GConsent to define the source of their concepts and relationships. By updating GDPRtEXT to use the updated ELI ontology, the interpretation of GDPR as a linked data resource can be provided using the authoritative URIs for use with the provided SKOS vocabulary.

Create vocabulary for expressing GDPR Compliance

The vocabularies associated with GDPR, including those presented in the state of the art in Chapter 3 and as contributions of the thesis in Chapter 5, address compliance by associating information with its requirements. This establishes the opportunity to create a vocabulary that represents compliance itself by describing the state of information in fulfilling requirements. Such a vocabulary would be of use to supervisory authorities as well as controllers and processors in generating documentation demonstrating the compliance of information as well as the degree to which it was fulfilled or achieved.

Expand GConsent to capture real-world interactions on the web

The aim of GConsent, as presented in this thesis, is to represent information about consent. While it is a comprehensive and detailed ontology compared to the state of the art, it currently is not sufficient to express the nuances and complexities of real-world interactions - such as those found in the consent mechanisms on websites. More specifically, it lacks a way to describe the intricate relationships of different organisations, including third parties, and the combined collection and dissemination of consent which happens via real-time bidding online. This can be remedied by incorporating legal opinions on online consent as they appear in the coming periods of time. Currently, GConsent is also being used in conjunction with the DPV to create an updated Consent Receipt [170] based on requirements of consent under GDPR.

Generate Assistive Systems for Compliance

The research presented in this thesis provides a technological base for modeling, querying, and validating information associated with GDPR compliance. In order for organisations to make use of this research - they need to express their use-case and internal activities using the developed ontologies. This presents an opportunity for tools and assistive technologies to be developed for helping organisations with the task of information gathering and documentation associated with GDPR compliance. These can be commercial products - similar

to existing ones - or a collaborative community effort that takes advantage of the interoperability provided by semantic web. With this background, following are three opportunities where this research can be applied.

Incorporate GDPRov and GConsent in the SPECIAL compliance checker

The compliance checker developed by the SPECIAL project [85] uses a semantic reasoner [61] with a controlled vocabulary consisting of personal data, processing, purpose, storage, and recipients expressed in OWL2. It is potentially possible to use the SPECIAL compliance checker to check the compliance of information defined using GDPRov and GConsent by modifying the checker to target these vocabularies or by alignment of SPECIAL vocabularies with GDPRov and GConsent. This would enable the work presented in this thesis to take advantage of large scale analysis and transparent log mechanisms provided by the SPECIAL architecture. Evaluation of the approach would be based on analysis of scalability and performance to ascertain extent of its benefit.

Privacy Policy annotation and automatic generation

A privacy policy fulfils the legal requirement for dissemination of information concerning the processing of personal data. Existing approaches for annotating privacy policies [164] do not take into account the semantics of associated information, nor effects of GDPR on privacy policy as a document. The argument for a privacy policy dataset specifically annotated for GDPR [166] consists of using concepts relevant to the legislation in the annotation process. This can be achieved through use of GDPRtEXT as a vocabulary of GDPR concepts. In addition, workflows represented using GDPRov provide the necessary information in order to generate a partial privacy policy, and can be used to automate generation of text by converting the processing workflows into natural language text. An early exploration of this work regarding annotation of privacy policy and personalising was presented in [45]. Providing privacy policies with machine-readable metadata would assist in the automated information extraction regarding processing activities and provide assistance to supervisory authorities and data subjects in evaluating an organisation's practices.

Design patterns for GDPR compliance

While there are verbose ontologies to represent information associated with compliance, their specific usage is dependant on the applied use-case. To facilitate adoption and usage, a library of design patterns can be created where each pattern is concerned with representing information associated with compliance for a specific concept or clause of the GDPR. For example: a design pattern representing periodic collection of GPS data from smartphone devices, which is linked with applicable clauses of GDPR as well as requirements or constraints it must fulfil in order to be compliant. Such design patterns can be used as the basis for assistive tools that generate and assess information for compliance.

7.5 Final Remarks

GDPR is the subject of scrutiny due to its impending interpretation by supervisory authorities and courts of law and the possibility of incurring large amount of fines. Consequently, there is significant interest in approaches associated with its compliance, particularly those that involve technological means as they promise algorithmic solutions that can be automated.

Technological solutions towards addressing compliance are dependant on the underlying information model, and have a range of approaches to choose from - as is evident in the state of the art regarding regulatory compliance. However, it can be argued that the law ultimately only deals with legal documentation where information is invariably linked with specific clauses of the law.

Within this context, the work presented in this thesis is useful for all involved stakeholders - controllers, processors, supervisory authorities, and data subjects - by enabling creation of tools and services to assist in the representation, querying, and validation of information. In particular, the thesis establishes advantages of using semantic web technologies and provides an argument towards their adoption in the regulatory compliance domain. Where use-cases and context differs, stakeholders now have the technological means towards establishing common patterns and tools beneficial to the larger community.

With an increased need and focus on the intersection of technology and privacy, approaches based on semantic web can foster transparency and accountability by enabling an open medium for knowledge interaction for all stakeholders. It is therefore the author's hope that this thesis and the work presented therein is of benefit to society for meeting the expectations demanded by privacy laws such as GDPR as well those arising from social obligations.

This page intentionally left blank.

References

[1] G. Greenleaf, "Global Data Privacy Laws 2019: 132 National Laws & Many Bills", *Privacy Laws & Business International Report*, pp. 14–18, 2019.

[2] W. Christl and S. Spiekermann, *Networks of Control: A Report on Corporate Surveillance, Digital Tracking, Big Data & Privacy*. Wien: Facultas, 2016, 165 pp., ISBN: 978-3-7089-1473-2.

[3] "Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)", *Official Journal of the European Union*, vol. L119, May 2016. [Online]. Available: http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L:2016:119:TOC.

[4] (2019). GDPR Enforcement Tracker - list of GDPR fines, [Online]. Available: http://www.enforcementtracker.com (visited on 09/15/2019).

[5] (Jan. 21, 2019). The CNIL's restricted committee imposes a financial penalty of 50 Million euros against GOOGLE LLC | CNIL, [Online]. Available: https://www.cnil.fr/en/cnils-restricted-committee-imposes-financial-penalty-50-million-euros-against-google-llc (visited on 01/21/2019).

[6] "Directive 95/46/EC of the European Parliament and of the Council on the Protection of Individuals with Regard to the Processing of Personal Data and on the Free Movement of Such Data", *Official Journal of the European Union*, vol. L281, pp. 31–50, Nov. 1995. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:31995L0046.

[7] K. McCullagh, O. Tambou, and S. Bourton, *National Adaptations of the GDPR*, 2019. [Online]. Available: https://wp.me/p6OBGR-3dP (visited on 04/09/2019).

[8] "Assembly Bill No. 375 Chapter 55: An act to add Title 1.81.5 (commencing with Section 1798.100) to Part 4 of Division 3 of the Civil Code, relating to privacy", *California State Legislature*, Jun. 29, 2018, Key: Assembly Bill No. 375. [Online]. Available: https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201720180AB375 (visited on 07/24/2019).

[9] D. Machuletz and R. Böhme, "Multiple Purposes, Multiple Problems: A User Study of Consent Dialogs after GDPR", Aug. 27, 2019. arXiv: 1908.10048 [cs]. [Online]. Available: http://arxiv.org/abs/1908.10048 (visited on 09/02/2019).

[10] C. Utz, M. Degeling, S. Fahl, F. Schaub, and T. Holz, "(Un)informed Consent: Studying GDPR Consent Notices in the Field", in *ACM SIGSAC Conference on Computer and Communications Security (CCS'19)*, London, United Kingdom, Nov. 11–15, 2019, p. 18.

[11] (Mar. 21, 2019). Opinion of Advocate General - Case C-673/17, [Online]. Available: http://curia.europa.eu/juris/document/document.jsf?docid=212023&mode=req&pageIndex=1&dir=&occ=first&part=1&text=&doclang=EN&cid=5704393 (visited on 03/29/2019).

[12] P. N. Otto and A. I. Anton, "Addressing Legal Requirements in Requirements Engineering", in *15th IEEE International Requirements Engineering Conference (RE 2007)*, IEEE, Oct. 2007, pp. 5–14, ISBN: 978-0-7695-2935-6. DOI: 10/d4rpf3. [Online]. Available: http://ieeexplore.ieee.org/document/4384161/ (visited on 02/13/2017).

[13] S. Sadiq, G. Governatori, and K. Namiri, "Modeling Control Objectives for Business Process Compliance", in *Business Process Management*, G. Alonso, P. Dadam, and M. Rosemann, Eds., ser. Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2007, pp. 149–164, ISBN: 978-3-540-75183-0.

[14] T. F. Gordon, G. Governatori, and A. Rotolo, "Rules and norms: Requirements for rule interchange languages in the legal domain", in *International Workshop on Rules and Rule Markup Languages for the Semantic Web*, Springer, 2009, pp. 282–296. DOI: 10/fwf8xf. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-642-04985-9_26 (visited on 02/13/2017).

[15] M. Fellmann and A. Zasada, "STATE-OF-THE-ART OF BUSINESS PROCESS COMPLIANCE APPROACHES", *Tel Aviv*, p. 18, 2014.

[16] O. Akhigbe, D. Amyot, and G. Richards, "Information Technology Artifacts in the Regulatory Compliance of Business Processes: A Meta-Analysis", in *E-Technologies*, M. Benyoucef, M. Weiss, and H. Mili, Eds., vol. 209, Cham: Springer International Publishing, 2015, pp. 89–104, ISBN: 978-3-319-17956-8 978-3-319-17957-5. DOI: 10.1007/978-3-319-17957-5_6. [Online]. Available: http://link.springer.com/10.1007/978-3-319-17957-5_6 (visited on 05/23/2019).

[17] A. Elgammal, O. Turetken, W.-J. van den Heuvel, and M. Papazoglou, "Formalizing and appling compliance patterns for business process compliance", *Software & Systems Modeling*, vol. 15, no. 1, pp. 119–146, Feb. 2016, ISSN: 1619-1366, 1619-1374. DOI: 10/gfzrgw. [Online]. Available: http://link.springer.com/10.1007/s10270-014-0395-3 (visited on 05/31/2019).

[18] S. Kirrane, A. Mileo, and S. Decker, "Access control and the Resource Description Framework: A survey", *Semantic Web*, vol. 8, no. 2, B. Cuenca Grau, Ed., pp. 311–352, Dec. 6, 2016, ISSN: 22104968, 15700844. DOI: 10/gfxvr7. [Online]. Available: http://www.medra.org/servlet/aliasResolver?alias=iospress&doi=10.3233/SW-160236 (visited on 04/18/2018).

[19] C. Bizer, T. Heath, and T. Berners-Lee, "Linked Data: The Story so Far", *Semantic Services, Interoperability and Web Applications: Emerging Concepts*, pp. 205–227, 2011. DOI: 10/dh8v52. [Online]. Available: https://www.igi-global.com/chapter/linked-data-story-far/55046 (visited on 09/17/2019).

[20] M. Palmirani, R. Sperberg, G. Vergottini, and F. Vitali, *Akoma Ntoso Version 1.0. Part 1: XML Vocabulary*, Aug. 29, 2018. [Online]. Available: http://docs.oasis-open.org/legaldocml/akn-core/v1.0/akn-core-v1.0-part1-vocabulary.html (visited on 09/17/2019).

[21] European Union, Publications Office, and ELI Task Force, *ELI Implementation Methodology: Good Practices and Guidelines.* Luxembourg: Publications Office, 2015, OCLC: 948769714, ISBN: 978-92-78-41354-5.

[22] M. van Opijnen, "European Case Law Identifier: Indispensable Asset for Legal Information Retrieval", *From Information to Knowledge – Online Access to Legal Information: Methodologies, Trends and Perspectives*, p. 12, Dec. 2011.

[23] (2015). Semantic Web - W3C, [Online]. Available: https://www.w3.org/standards/semanticweb/ (visited on 09/17/2019).

[24] (Jun. 24, 2014). RDF 1.1 Primer, [Online]. Available: https://www.w3.org/TR/rdf11-primer/ (visited on 08/11/2019).

[25] C. M. d. O. Rodrigues, F. L. G. de Freitas, E. F. S. Barreiros, R. R. de Azevedo, and A. T. de Almeida Filho, "Legal ontologies over time: A systematic mapping study", *Expert Systems with Applications*, vol. 130, pp. 12–30, Sep. 15, 2019, ISSN: 0957-4174. DOI: 10/gf223z. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417419302398 (visited on 05/23/2019).

[26] (2019). Scalable Policy-aware Linked Data Architecture For Privacy, Transparency and Compliance (SPECIAL), [Online]. Available: https://www.specialprivacy.eu/ (visited on 09/17/2019).

[27] (2019). MIREL - MIning and REasoning with Legal texts, [Online]. Available: http://www.mirelproject.eu/ (visited on 09/17/2019).

[28] (2019). DAta Protection REgulation COmpliance (DAPRECO), [Online]. Available: https://www.fnr.lu/projects/data-protection-regulation-compliance/ (visited on 09/17/2019).

[29] (2019). BPR4GDPR, [Online]. Available: http://www.bpr4gdpr.eu/ (visited on 08/27/2019).

[30] (2019). RestAssured, [Online]. Available: https://restassuredh2020.eu/ (visited on 08/27/2019).

[31] (Dec. 11, 2012). OWL 2, [Online]. Available: https://www.w3.org/TR/owl2-overview/ (visited on 08/11/2019).

[32] (2013). SPARQL 1.1 Query Language, [Online]. Available: https://www.w3.org/TR/sparql11-query/ (visited on 04/30/2019).

[33] H. Knublauch and D. Kontokostas. (Jul. 2017). Shapes Constraint Language (SHACL), [Online]. Available: https://www.w3.org/TR/shacl/ (visited on 09/19/2018).

[34] M. C. Suárez-Figueroa, A. Gómez-Pérez, and M. Fernández-López, "The NeOn Methodology for Ontology Engineering", in *Ontology Engineering in a Networked World*, M. C. Suárez-Figueroa, A. Gómez-Pérez, E. Motta, and A. Gangemi, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 9–34, ISBN: 978-3-642-24793-4 978-3-642-24794-1. DOI: 10.1007/978-3-642-24794-1_2. [Online]. Available: http://link.springer.com/10.1007/978-3-642-24794-1_2 (visited on 11/27/2017).

[35] N. F. Noy, D. L. McGuinness, *et al.*, *Ontology development 101: A guide to creating your first ontology*, 2001.

[36] A. De Nicola and M. Missikoff, "A lightweight methodology for rapid ontology engineering", *Communications of the ACM*, vol. 59, no. 3, pp. 79–86, Feb. 25, 2016, ISSN: 00010782. DOI: 10/gftgpt. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2897191.2818359 (visited on 01/16/2019).

[37] D. Garijo, "WIDOCO: A wizard for documenting ontologies", in *International Semantic Web Conference*, Springer, 2017, pp. 94–102. DOI: 10/gfxvtk.

[38] M. Poveda-Villalón, A. Gómez-Pérez, and M. C. Suárez-Figueroa, "OOPS! (OntOlogy Pitfall Scanner!): An On-line Tool for Ontology Evaluation", *International Journal on Semantic Web and Information Systems (IJSWIS)*, vol. 10, no. 2, pp. 7–34, Apr. 1, 2014, ISSN: 1552-6283 DOI: 10.4018/ijswis.2014040102. DOI: 10/f6qxmv. [Online]. Available: https://www.igi-global.com/article/oops-ontology-pitfall-scanner/116450 (visited on 03/25/2019).

[39] F. Thomas, D. John, P. Roberto, M. Carmen, and P. Marco, "The European Legislation Identifier", *Frontiers in Artificial Intelligence and Applications*, pp. 137–148, 2019, ISSN: 0922-6389. DOI: 10/gf7fbr. [Online]. Available: http://www.medra.org/servlet/aliasResolver?alias=iospressISBN&isbn=978-1-61499-984-3&spage=137&doi=10.3233/FAIA190016 (visited on 09/02/2019).

[40] S. Agarwal, S. Steyskal, F. Antunovic, and S. Kirrane, "Legislative Compliance Assessment: Framework, Model and GDPR Instantiation", in *Privacy Technologies and Policy*, M. Medina, A. Mitrakas, K. Rannenberg, E. Schweighofer, and N. Tsouroulas, Eds., ser. Lecture Notes in Computer Science, Springer International Publishing, 2018, pp. 131–149, ISBN: 978-3-030-02547-2.

[41] M. Palmirani, M. Martoni, A. Rossi, C. Bartolini, and L. Robaldo, "PrOnto: Privacy Ontology for Legal Compliance", in *Proceedings of the 18th European Conference on Digital Government ECDG 2018*, 2018, p. 10.

[42] "Council conclusions inviting the introduction of the European Legislation Identifier (ELI)", *Official Journal of the European Union*, vol. 325, no. 3, pp. 3–11, Oct. 26, 2012. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52012XG1026(01) (visited on 09/17/2019).

[43] V. Leone, L. Di Caro, and S. Villata, "Taking stock of legal ontologies: A feature-based comparative analysis", *Artificial Intelligence and Law*, Jun. 13, 2019, ISSN: 1572-8382. DOI: 10/gf3z84. [Online]. Available: https://doi.org/10.1007/s10506-019-09252-1 (visited on 06/18/2019).

[44] H. J. Pandit, D. O'Sullivan, and D. Lewis, "An Ontology Design Pattern for Describing Personal Data in Privacy Policies", in *Proceedings of the 9th Workshop on Ontology Design and Patterns (WOP 2018) Co-Located with 17th International Semantic Web Conference (ISWC 2018)*, Monterey, California, USA, 2018. [Online]. Available: http://ceur-ws.org/Vol-2195/pattern_paper_3.pdf.

[45] H. J. Pandit, D. O'Sullivan, and D. Lewis, "Personalised Privacy Policies", in *New Trends in Databases and Information Systems*, A. Benczúr, B. Thalheim, T. Horváth, S. Chiusano, T. Cerquitelli, C. Sidló, and P. Z. Revesz, Eds., ser. Communications in Computer and Information Science, Springer International Publishing, 2018, pp. 127–137, ISBN: 978-3-030-00063-9. DOI: 10/gfxgwn.

[46] H. J. Pandit, D. O'Sullivan, and D. Lewis, "GDPR-driven Change Detection in Consent and Activity Metadata", in *Joint Proceedings of the 4th Workshop on Managing the Evolution and Preservation of the Data Web (MEPDaW), the 2nd Workshop on Semantic Web Solutions for Large-Scale Biomedical Data Analytics (SeWeBMeDA), and the Workshop on Semantic Web of Things for Industry 4.0 (SWeTI) Co-Located with 15th European Semantic Web Conference (ESWC 2018)*, Heraklion, Crete, Greece, 2018, p. 5. [Online]. Available: http://ceur-ws.org/Vol-2112/mepdaw_paper_2.pdf.

[47] T. Lebo, S. Sahoo, D. McGuinness, K. Belhajjame, J. Cheney, D. Corsar, D. Garijo, S. Soiland-Reyes, S. Zednik, and J. Zhao. (2013). PROV-O: The PROV Ontology.

[48] D. Garijo and Y. Gil. (Mar. 12, 2014). The P-PLAN Ontology, [Online]. Available: http://vocab.linkeddata.es/p-plan/ (visited on 09/19/2018).

[49] T. Pasquier, J. Singh, J. Powles, D. Eyers, M. Seltzer, and J. Bacon, "Data provenance to audit compliance with privacy policy in the Internet of Things", *Personal and Ubiquitous Computing*, vol. 22, no. 2, pp. 333–344, Apr. 1, 2018, ISSN: 1617-4909, 1617-4917. DOI: 10/gdcvmb. [Online]. Available: https://link.springer.com/article/10.1007/s00779-017-1067-4 (visited on 05/02/2018).

[50] B. E. Ujcich, A. Bates, and W. H. Sanders, "A Provenance Model for the European Union General Data Protection Regulation", in *Provenance and Annotation of Data and Processes*, K. Belhajjame, A. Gehani, and P. Alper, Eds., vol. 11017, Cham: Springer International Publishing, 2018, pp. 45–57, ISBN: 978-3-319-98378-3 978-3-319-98379-0. DOI: 10.1007/978-3-319-98379-0_4. [Online]. Available: http://link.springer.com/10.1007/978-3-319-98379-0_4 (visited on 09/10/2018).

[51] P. A. Bonatti, W. Dullaert, J. D. Fernandez, S. Kirrane, U. Milosevic, and A. Polleres, *The SPECIAL Policy Log Vocabulary V0.5*, 2018. [Online]. Available: https://aic.ai.wu.ac.at/qadlod/policyLog/.

[52] W. Dullaert, U. Milosevic, J. Langens, A. S'Jongers, N. Szepes, V. Goossens, N. Rudavsky-Brody, W. Delabastita, S. Kirrane, and J. D. Fernandez, *D3.4 Transparency & Compliance Release*, Jan. 31, 2019. [Online]. Available: https://www.specialprivacy.eu/images/documents/SPECIAL_D34_M25_V10.pdf (visited on 05/31/2019).

[53] E. Politou, E. Alepis, and C. Patsakis, "Forgetting personal data and revoking consent under the GDPR: Challenges and proposed solutions", *Journal of Cybersecurity*, vol. 4, no. 1, Jan. 1, 2018, ISSN: 2057-2085. DOI: 10/gfsqrg. [Online]. Available: https://academic.oup.com/cybersecurity/article/4/1/tyy001/4954056 (visited on 04/01/2019).

[54] A. 2. D. P. W. Party, *Guidelines on Consent under Regulation 2016/679 (wp259rev.01)*, Apr. 10, 2018. [Online]. Available: https://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=623051 (visited on 12/06/2018).

[55] M. Palmirani, M. Martoni, A. Rossi, C. Bartolini, and L. Robaldo, "PrOnto: Privacy Ontology for Legal Reasoning", in *Electronic Government and the Information Systems Perspective*, A. Kő and E. Francesconi, Eds., ser. Lecture Notes in Computer Science, Springer International Publishing, 2018, pp. 139–152, ISBN: 978-3-319-98349-3.

[56] *GDPR Readiness Checklist Template for SMEs*, Dec. 2017.

[57] H. J. Pandit, D. O'Sullivan, and D. Lewis, "Queryable Provenance Metadata For GDPR Compliance", in *Procedia Computer Science*, ser. Proceedings of the 14th International Conference on Semantic Systems 10th – 13th of September 2018 Vienna, Austria, vol. 137, Jan. 1, 2018, pp. 262–268. DOI: 10/gfdc6r. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1877050918316314 (visited on 10/18/2018).

[58] H. J. Pandit, D. O'Sullivan, and D. Lewis, "Towards Knowledge-based Systems for GDPR Compliance", in *Proceedings of the Joint Proceedings of the International Workshops on Contextualized Knowledge Graphs, and Semantic Statistics (CKGSemStats)*, Monterey, California, USA, 2018. [Online]. Available: http://ceur-ws.org/Vol-2317/article-09.pdf.

[59] ——, "Exploring GDPR Compliance Over Provenance Graphs Using SHACL", in *Proceedings of the Posters and Demos Track of the 14th International Conference on Semantic Systems Co-Located with the 14th International Conference on Semantic Systems (SEMANTiCS 2018)*, Vienna, Austria, 2018. [Online]. Available: http://ceur-ws.org/Vol-2198/paper_120.pdf.

[60] H. J. Pandit, D. O'Sullivan, and D. Lewis, "Test-driven Approach Towards GDPR Compliance", in *15th International Conference on Semantic Systems (SEMANTiCS2019)*, Karlsruhe, Germany, 2019.

[61] P. A. Bonatti, "Fast Compliance Checking in an OWL2 Fragment", in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, Stockholm, Sweden: International Joint Conferences on Artificial Intelligence Organization, Jul. 2018, pp. 1746–1752, ISBN: 978-0-9992411-2-7. DOI: 10/gfxvsm. [Online]. Available: https://www.ijcai.org/proceedings/2018/241 (visited on 08/28/2018).

[62] M. D. Vos, S. Kirrane, J. Padget, and K. Satoh, "ODRL policy modelling and compliance checking", in *3rd International Joint Conference on Rules and Reasoning (RuleML+RR 2019)*, Bolzano, Italy, Sep. 16–19, 2019, p. 16.

[63] P. Monica and G. Guido, "Modelling Legal Knowledge for GDPR Compliance Checking", *Frontiers in Artificial Intelligence and Applications*, pp. 101–110, 2018, ISSN: 0922-6389. DOI: 10/gfr9qc. [Online]. Available: http://www.medra.org/servlet/aliasResolver?alias=iospressISBN&isbn=978-1-61499-934-8&spage=101&doi=10.3233/978-1-61499-935-5-101 (visited on 12/24/2018).

[64] A. S. M. Mehr, "Compliance to data protection and purpose control using process mining technique", in *Proceedings of the Dissertation Award, Doctoral Consortium, and Demonstration Track at BPM 2019 Co-Located with 17th International Conference on Business Process Management (BPM 2019)*, Vienna, Austria, Sep. 1–6, 2019, p. 6.

[65] S. Lieber, "Policy-compliant Data Processing: RDF-based Restrictions for Data-protection", in *Doctoral Track - 18th International Semantic Web Conference (ISWC)*, Auckland, New Zealand, 2019, p. 12.

[66] H. J. Pandit and D. Lewis, "Modelling Provenance for GDPR Compliance using Linked Open Data Vocabularies", in *Proceedings of the 5th Workshop on Society, Privacy and the Semantic Web - Policy and Technology (PrivOn2017) (PrivOn)*, 2017. [Online]. Available: http://ceur-ws.org/Vol-1951/PrivOn2017_paper_6.pdf.

[67] H. J. Pandit, C. Debruyne, D. O'Sullivan, and D. Lewis, "An Exploration of Data Interoperability for GDPR", *International Journal of Standardization Research (IJSR)*, vol. 16, no. 1, pp. 1–21, Jan. 1, 2018, ISSN: 2470-8542 DOI: 10.4018/IJSR.2018010101. DOI: 10/gfsn52. [Online]. Available: https://www.igi-global.com/article/an-exploration-of-data-interoperability-for-gdpr/218518 (visited on 01/04/2019).

[68] H. J. Pandit, C. Debruyne, D. O'Sullivan, and D. Lewis. (2020). Standardisation, Data Interoperability, and GDPR.

[69] H. J. Pandit and A. Polleres. (Jul. 26, 2019). DPV, [Online]. Available: https://www.w3.org/ns/dpv (visited on 08/11/2019).

[70] H. J. Pandit, A. Polleres, B. Bos, R. Brennan, B. Bruegger, F. J. Ekaputra, J. D. Fernandez, R. G. Hamed, E. Kiesling, M. Lizar, E. Schlehahn, S. Steyskal, R. Wenning, and M. Kurze, *D6.5 Final Report of the Community Group*, Jul. 31, 2019. [Online]. Available: https://www.specialprivacy.eu/images/documents/SPECIAL_D65_M30_V10.pdf (visited on 08/21/2019).

[71] H. J. Pandit, C. Debruyne, D. O'Sullivan, and D. Lewis, "GConsent - A Consent Ontology Based on the GDPR", in *The Semantic Web*, P. Hitzler, M. Fernández, K. Janowicz, A. Zaveri, A. J. Gray, V. Lopez, A. Haller, and K. Hammar, Eds., ser. Lecture Notes in Computer Science, Springer International Publishing, 2019, pp. 270–282, ISBN: 978-3-030-21348-0. [Online]. Available: https://w3id.org/GConsent.

[72] H. J. Pandit, K. Fatema, D. O'Sullivan, and D. Lewis, "GDPRtEXT - GDPR as a Linked Data Resource", in *The Semantic Web - European Semantic Web Conference*, ser. Lecture Notes in Computer Science, Springer, Cham, Jun. 3, 2018, pp. 481–495, ISBN: 978-3-319-93416-7 978-3-319-93417-4. DOI: 10/c3n4. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-93417-4_31 (visited on 06/17/2018).

[73] K. Fatema, E. Hadziselimovic, H. J. Pandit, C. Debruyne, D. Lewis, and D. O'Sullivan, "Compliance through Informed Consent: Semantic Based Consent Permission and Data Management Model", in *Proceedings of the 5th Workshop on Society, Privacy and the Semantic Web - Policy and Technology (PrivOn2017) (PrivOn)*, 2017. [Online]. Available: http://ceur-ws.org/Vol-1951/PrivOn2017_paper_5.pdf.

[74] E. Hadziselimovic, K. Fatema, H. J. Pandit, and D. Lewis, "Linked Data Contracts to Support Data Protection and Data Ethics in the Sharing of Scientific Data", in *Proceedings of the First Workshop on Enabling Open Semantic Science (SemSci)*, 2017, pp. 55–62. [Online]. Available: http://ceur-ws.org/Vol-1931/paper-08.pdf.

[75] H. J. Pandit, D. O'Sullivan, and D. Lewis, "GDPR Data Interoperability Model", in *23rd EURAS Annual Standardisation Conference*, Dublin, Ireland, 2018. [Online]. Available: http://openscience.adaptcentre.ie/pb/EURAS2018/.

[76] C. Debruyne, H. J. Pandit, D. Lewis, and D. O'Sullivan, "Towards Generating Policy-Compliant Datasets", in *2019 IEEE 13th International Conference on Semantic Computing (ICSC)*, Jan. 2019, pp. 199–203. DOI: 10/gfxgwx.

[77] H. J. Pandit, D. O'Sullivan, and D. Lewis, "Extracting Provenance Metadata from Privacy Policies", in *Provenance and Annotation of Data and Processes*, K. Belhajjame, A. Gehani, and P. Alper, Eds., ser. Lecture Notes in Computer Science, Springer International Publishing, 2018, pp. 262–265, ISBN: 978-3-319-98379-0. DOI: 10/gfxgwm.

[78] H. J. Pandit, A. Polleres, B. Bos, R. Brennan, B. Bruegger, F. J. Ekaputra, J. D. Fernández, R. G. Hamed, M. Lizar, E. Schlehahn, S. Steyskal, and R. Wenning, "Creating A Vocabulary for Data Privacy", in *The 18th International Conference on Ontologies, DataBases, and Applications of Semantics (ODBASE2019)*, Rhodes, Greece, 2019, p. 17.

[79] A. Marini, A. Kateifides, J. Bates, G. Zanfir-Fortuna, M. Bae, S. Gray, and G. Sen, *GDPR CCPA Comparison Guide*, Nov. 2018. [Online]. Available: https://fpf.org/wp-content/uploads/2018/11/GDPR_CCPA_Comparison-Guide.pdf (visited on 03/31/2019).

[80] *GDPR Compliance Checklist*, 2019.

[81] *Privacy and Security by Design Methodology Handbook*, 2015. [Online]. Available: http://pripareproject.eu/wp-content/uploads/2013/11/PRIPARE-Methodology-Handbook-Final-Feb-24-2016.pdf (visited on 08/27/2019).

[82] T. Pellegrini, A. Schönhofer, S. Kirrane, A. Fensel, O. Panasiuk, V. Mireles, T. Thurner, A. Polleres, and M. Dörfler, "A genealogy and classification of rights expression languages-preliminary results", in *Data Protection/LegalTech-Proceedings of the 21st International Legal Informatics Symposium IRIS*, 2018, pp. 243–250.

[83] J. van de Ven and F. Dylla, "Qualitative Privacy Description Language", in *Privacy Technologies and Policy: Annual Privacy Forum 2016*, S. Schiffner, J. Serna, D. Ikonomou, and K. Rannenberg, Eds., ser. Lecture Notes in Computer Science, Springer International Publishing, 2016, pp. 171–189, ISBN: 978-3-319-44760-5. DOI: 10/gf6mmb.

[84] M. Gharib, P. Giorgini, and J. Mylopoulos, "Ontologies for Privacy Requirements Engineering: A Systematic Literature Review", Nov. 30, 2016. arXiv: 1611.10097. [Online]. Available: http://arxiv.org/abs/1611.10097 (visited on 04/01/2019).

[85] S. Kirrane, J. D. Fernández, W. Dullaert, U. Milosevic, A. Polleres, P. Bonatti, R. Wenning, O. Drozd, and P. Raschke, "A Scalable Consent, Transparency and Compliance Architecture", in *Proceedings of the Posters and Demos Track of the Extended Semantic Web Conference (ESWC 2018)*, 2018. DOI: 10/gfxvsf.

[86] M. Palmirani, G. Governatori, A. Rotolo, S. Tabet, H. Boley, and A. Paschke, "LegalRuleML: XML-Based Rules and Norms", in *Rule-Based Modeling and Computing on the Semantic Web*, F. Olken, M. Palmirani, and D. Sottara, Eds., ser. Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2011, pp. 298–312, ISBN: 978-3-642-24908-2.

[87] G. Governatori, M. Hashmi, H.-P. Lam, S. Villata, and M. Palmirani, "Semantic Business Process Regulatory Compliance Checking Using LegalRuleML", in *Knowledge Engineering and Knowledge Management*, E. Blomqvist, P. Ciancarini, F. Poggi, and F. Vitali, Eds., ser. Lecture Notes in Computer Science, Springer International Publishing, 2016, pp. 746–761, ISBN: 978-3-319-49004-5.

[88] N. Syed Abdullah, S. Sadiq, and M. Indulska, "A Compliance Management Ontology: Developing Shared Understanding through Models", in *Advanced Information Systems Engineering*, J. Ralyté, X. Franch, S. Brinkkemper, and S. Wrycza, Eds., ser. Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2012, pp. 429–444, ISBN: 978-3-642-31095-9.

[89] P. Casanovas, J. González-Conejero, and L. de Koker, "Legal Compliance by Design (LCbD) and through Design (LCtD): Preliminary Survey", in *Proceedings of the 1st Workshop on Technologies for Regulatory Compliance Co-Located with the 30th International Conference on Legal Knowledge and Information Systems (JURIX 2017), Luxembourg, December 13, 2017.*, 2017, pp. 33–49. [Online]. Available: http://ceur-ws.org/Vol-2049/05paper.pdf.

[90] O. Drozd and S. Kirrane, "Consent Comprehension Made Easy Demo", in *19th Privacy Enhancing Technologies Symposium (PETS)*, 2019, p. 1.

[91] ——, "I Agree: Customize Your Personal Data Processing with the CoRe User Interface", in *Trust, Privacy and Security in Digital Business*, S. Gritzalis, E. R. Weippl, S. K. Katsikas, G. Anderst-Kotsis, A. M. Tjoa, and I. Khalil, Eds., vol. 11711, Cham: Springer International Publishing, 2019, pp. 17–32, ISBN: 978-3-030-27812-0 978-3-030-27813-7. DOI: 10.1007/978-3-030-27813-7_2. [Online]. Available: http://link.springer.com/10.1007/978-3-030-27813-7_2 (visited on 08/21/2019).

[92] P. Westphal, J. D. Fernandez, and S. Kirrane, "SPIRIT: A Semantic Transparency and Compliance Stack", in *Proceedings of the 14th International Conference on Semantic Systems (SEMANTiCS)*, 2018, p. 4.

[93] J. D. Fernandez, M. Sabou, E. Kiesling, F. J. Ekaputra, A. Azzam, and R. Wenning, "User Consent Modeling for Ensuring Transparency and Compliance in Smart Cities", *Personal and Ubiquitous Computing Journal*, p. 34, 2019.

[94] P. A. Bonatti, B. Bos, S. Decker, J. D. Fernandez, V. Peristeras, A. Polleres, and R. Wenning, "Data Privacy Vocabularies and Controls: Semantic Web for Transparency and Privacy", in *Proceedings of the Workshop on Semantic Web for Social Good Co-Located with 17th International Semantic Web Conference (ISWC 2018)*, Monterey, California, USA, 2018, p. 4. [Online]. Available: http://ceur-ws.org/Vol-2182/paper_3.pdf.

[95] P. A. Bonatti, S. Kirrane, I. M. Petrova, L. Sauro, and E. Schlehahn, *The SPECIAL Usage Policy Language V0.1*, 2018. [Online]. Available: http://purl.org/specialprivacy/policylanguage.

[96] P. A. Bonatti, I. M. Petrova, and L. Sauro, "A richer policy language for GDPR compliance", in *Proceedings of the 32nd International Workshop on Description Logics*, Oslo, Norway, Jun. 18–21, 2019, p. 12.

[97] P. A. Bonatti, J. Colbeck, F. D. Meersman, R. Jacob, S. Kirrane, M. Kurze, M. Piekarska, R. Wenning, B. Whittam-Smith, H. Zwingelberg, E. Schlehahn, and L. Sauro, *D1.5 Use case scenarios V2*, Feb. 28, 2018. [Online]. Available: https://www.specialprivacy.eu/images/documents/SPECIAL_D1.5_M14_V1.0.pdf (visited on 10/12/2018).

[98] E. Schlehahn and R. Wenning, *D1.6 Legal requirements for a privacy-enhancing Big Data V2*, Apr. 28, 2018. [Online]. Available: https://www.specialprivacy.eu/images/documents/SPECIAL_D1.6_M15_V1.0.pdf (visited on 06/06/2018).

[99] P. A. Bonatti, S. Kirrane, I. M. Petrova, L. Sauro, and E. Schlehahn, *D2.5 Policy Language V2*, Dec. 31, 2018. [Online]. Available: https://www.specialprivacy.eu/images/documents/SPECIAL_D25_M21_V10.pdf (visited on 05/31/2019).

[100] P. A. Bonatti, S. Kirrane, I. M. Petrova, L. Sauro, C. Kerschbaum, and E. Pirkova, *D2.6 Formal representation of the legislation V2*, Dec. 31, 2018. [Online]. Available: https://www.specialprivacy.eu/images/documents/SPECIAL_D26_M21_V10.pdf (visited on 05/31/2019).

[101] S. Kirrane, U. Milosevic, J. D. Fernandez, A. Polleres, and J. Langens, *D2.7 Transparency Framework V2*, Nov. 30, 2018. [Online]. Available: https://www.specialprivacy.eu/images/documents/SPECIAL_D27_M23_V10.pdf (visited on 05/31/2019).

[102] R. Wenning and S. Kirrane, "Compliance Using Metadata", in *Semantic Applications: Methodology, Technology, Corporate Use*, T. Hoppe, B. Humm, and A. Reibold, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2018, pp. 31–45, ISBN: 978-3-662-55433-3. DOI: 10.1007/978-3-662-55433-3_3. [Online]. Available: https://doi.org/10.1007/978-3-662-55433-3_3 (visited on 05/31/2019).

[103] S. Agarwal, S. Kirrane, and F. Antunovic, *D5.5 The GDPR Compliance Tool*, Jun. 16, 2017. [Online]. Available: https://cordis.europa.eu/docs/projects/cnect/2/612052/080/deliverables/001-D55v4GDPR.pdf (visited on 09/26/2018).

[104] J. D. Fernández, F. J. Ekaputra, P. Ruswono, E. Kiesling, and A. Azzam, "Privacy-aware Linked Widgets", in *1st Workshop on Fairness, Accountability, Transparency, Ethics, and Society on the Web. In Conjunction with The Web Conference 2019*, 2019, p. 8. DOI: 10/gf2599.

[105] *D6.1 Privacy policy formalization*, 2018. [Online]. Available: http://cityspin.net/wp-content/uploads/2017/10/D6.1-Privacy-policy-formalization.pdf (visited on 10/08/2019).

[106] *D6.3 Transparency framework*, 2019. [Online]. Available: http://cityspin.net/wp-content/uploads/2017/10/D6.3-Transparency-framework-1.pdf (visited on 10/08/2019).

[107] P. Raschke, A. Küpper, O. Drozd, and S. Kirrane, "Designing a gdpr-compliant and usable privacy dashboard", in *IFIP International Summer School on Privacy and Identity Management*, Springer, 2017, pp. 221–236.

[108] P. Raschke, O. Drozd, B. Bos, R. Jacob, and B. Whittamsmith, *D4.3 Transparency dashboard and control panel release V2*, 2019.

[109] U. Milošević, P. Raschke, O. Drozd, S. Kirrane, F. D. Meersman, and R. Jacob, *D4.4 Usability testing report V2*, Mar. 29, 2019.

[110] F. Gandon, G. Governatori, and S. Villata, "Normative Requirements as Linked Data", in *30th International Conference on Legal Knowledge and Information Systems (JURIX)*, Luxembourg, Dec. 2017, p. 11.

[111] M. Teruel, C. Cardellino, F. Cardellino, L. A. Alemany, and S. Villata, "Legal text processing within the MIREL project", in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[112] P. Monica, M. Michele, R. Arianna, B. Cesare, and R. Livio, "Legal Ontology for Modelling GDPR Concepts and Norms", *Frontiers in Artificial Intelligence and Applications*, pp. 91–100, 2018, ISSN: 0922-6389. DOI: 10/gfr9qd. [Online]. Available: http://www.medra.org/servlet/aliasResolver?alias=iospressISBN&isbn=978-1-61499-934-8&spage=91&doi=10.3233/978-1-61499-935-5-91 (visited on 12/24/2018).

[113] R. Arianna and P. Monica, "DaPIS: An Ontology-Based Data Protection Icon Set", *Frontiers in Artificial Intelligence and Applications*, pp. 181–195, 2019, ISSN: 0922-6389. DOI: 10/gf7fbn. [Online]. Available: http://www.medra.org/servlet/aliasResolver?alias=iospressISBN&isbn=978-1-61499-984-3&spage=181&doi=10.3233/FAIA190020 (visited on 09/02/2019).

[114] C. Bartolini, G. Lenzini, and L. Robaldo, "Towards legal compliance by correlating Standards and Laws with a semi-automated methodology", in *Proceedings of the 28 Benelux Conference on Artificial Intelligence (BNAIC)*, 2016. [Online]. Available: http://orbilu.uni.lu/handle/10993/28957 (visited on 09/13/2017).

[115] C. Bartolini, R. Muthuri, and C. Santos, "Using Ontologies to Model Data Protection Requirements in Workflows", in *New Frontiers in Artificial Intelligence*, M. Otake, S. Kurahashi, Y. Ota, K. Satoh, and D. Bekki, Eds., vol. 10091, Cham: Springer International Publishing, 2017, pp. 233–248, ISBN: 978-3-319-50952-5 978-3-319-50953-2. DOI: 10.1007/978-3-319-50953-2_17. [Online]. Available: http://link.springer.com/10.1007/978-3-319-50953-2_17 (visited on 10/08/2019).

[116] C. Bartolini, G. Lenzini, and C. Santos, "A Legal Validation of a Formal Representation of GDPR Articles", in *Proceedings of the 2nd Workshop on Technologies for Regulatory Compliance Co-Located with the 31st International Conference on Legal Knowledge and Information Systems (JURIX 2018)*, Groningen, Netherlands, 2018, p. 14.

[117] C. Bartolini, A. Calabró, and E. Marchetti, "Enhancing Business Process Modelling with Data Protection Compliance: An Ontology-based Proposal:" in *Proceedings of the 5th International Conference on Information Systems Security and Privacy*, Prague, Czech Republic: SCITEPRESS - Science and Technology Publications, 2019, pp. 421–428, ISBN: 978-989-758-359-9. DOI: 10/gf3czj. [Online]. Available: http://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0007392304210428 (visited on 05/31/2019).

[118] C. Bartolini, G. Lenzini, and C. Santos, "An Agile Approach to Validate a Formal Representation of the GDPR", *New Frontiers in Artificial Intelligence*, p. 16, 2019.

[119] R. Hoekstra, J. Breuker, M. Di Bello, A. Boer, *et al.*, "The LKIF Core Ontology of Basic Legal Concepts.", *LOAIT*, vol. 321, pp. 43–63, 2007.

[120] S. Peroni, D. Shotton, and F. Vitali, "Tools for the Automatic Generation of Ontology Documentation: A Task-Based Evaluation", *International Journal on Semantic Web and Information Systems (IJSWIS)*, vol. 9, no. 1, pp. 21–44, Jan. 1, 2013, ISSN: 1552-6283 DOI: 10.4018/jswis.2013010102. DOI: 10/f47ncz. [Online]. Available: https://www.igi-global.com/article/tools-automatic-generation-ontology-documentation/77823 (visited on 10/08/2019).

[121] L. Robaldo and X. Sun, "Reified Input/Output logic: Combining Input/Output logic and Reification to represent norms coming from existing legislation", *Journal of Logic and Computation*, vol. 27, no. 8, pp. 2471–2503, 2017. DOI: 10/gf9jmn.

[122] C. Cardellino, M. Teruel, L. A. Alemany, and S. Villata, "Legal NERC with ontologies, Wikipedia and curriculum learning", in *15th European Chapter of the Association for Computational Linguistics (EACL 2017)*, Valencia, Spain, 2017, pp. 254–259. DOI: 10/gf6rvp. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01572444 (visited on 08/21/2019).

[123] L. Robaldo, *D2.4 Ontology population: Connecting legal text toontology concepts and instances*, Dec. 20, 2017. [Online]. Available: http://www.mirelproject.eu/publications/D2.4.pdf (visited on 08/21/2019).

[124] G. Lioudakis and D. Cascone, *D3.1 Compliance Ontology*, Feb. 28, 2019.

[125] N. Dellas, *D2.3 Initial Specification of BPR4GDPR architecture*, Feb. 28, 2019.

[126] L. Elluri and K. P. Joshi, "A Knowledge Representation of Cloud Data Controls for EU GDPR Compliance", in *2018 IEEE World Congress on Services, SERVICES 2018, San Francisco, CA, USA, July 2-7, 2018*, 2018, pp. 45–46. DOI: 10/gft38j. [Online]. Available: https://doi.org/10.1109/SERVICES.2018.00036.

[127] L. Elluri, A. Nagar, and K. P. Joshi, "An Integrated Knowledge Graph to Automate GDPR and PCI DSS Compliance", in *2018 IEEE International Conference on Big Data (Big Data)*, Dec. 2018, pp. 1266–1271. DOI: 10/gf3cx9.

[128] K. P. Joshi and A. Banerjee, "Automating Privacy Compliance Using Policy Integrated Blockchain", *Cryptography*, vol. 3, no. 1, p. 7, Mar. 2019. DOI: 10/gf6rvc. [Online]. Available: https://www.mdpi.com/2410-387X/3/1/7 (visited on 08/21/2019).

[129] M. Winckler, L. Goncalves, O. Nicolas, F. Biennier, H. Benfenatki, T. Despeyroux, N. Alaya, A. Deslée, M. F. Diallo, I. Collin-Lachaud, G. Ubersfeld, and C. Cianchi, "Personal Information Controller Service (PICS)", in *Web Engineering*, M. Bakaev, F. Frasincar, and I.-Y. Ko, Eds., ser. Lecture Notes in Computer Science, Springer International Publishing, 2019, pp. 530–533, ISBN: 978-3-030-19274-7.

[130] H. Benfenatki, F. Biennier, M. Winckler, L. Goncalves, O. Nicolas, and Z. Saoud, "Towards a User Centric Personal Data Protection Framework", in *ACM CHI Conference on Human Factors in Computing Systems - GDPR Workshop*, 2018, p. 6.

[131] K. Rantos, G. Drosatos, K. Demertzis, C. Ilioudis, A. Papanikolaou, and A. Kritsas, "ADvoCATE: A Consent Management Platform for Personal Data Processing in the IoT Using Blockchain Technology", in *Innovative Security Solutions for Information Technology and Communications*, J.-L. Lanet and C. Toma, Eds., ser. Lecture Notes in Computer Science, Springer International Publishing, 2019, pp. 300–313, ISBN: 978-3-030-12942-2.

[132] C. Bartolini, R. Muthuri, and S. Cristiana, "Using ontologies to model data protection requirements in workflows", in *JSAI International Symposium on Artificial Intelligence*, 2015. [Online]. Available: http://orbilu.uni.lu/handle/10993/22383 (visited on 02/13/2017).

[133] M. Geko and S. Tjoa, "An Ontology Capturing the Interdependence of the General Data Protection Regulation (GDPR) and Information Security", in *Proceedings of the Central European Cybersecurity Conference 2018 on - CECC 2018*, Ljubljana, Slovenia: ACM Press, 2018, pp. 1–6, ISBN: 978-1-4503-6515-4. DOI: 10/gfxqw4. [Online]. Available: http://dl.acm.org/citation.cfm?doid=3277570.3277590 (visited on 04/01/2019).

[134] A. Gerl, N. Bennani, H. Kosch, and L. Brunie, "LPL, Towards a GDPR-Compliant Privacy Language: Formal Definition and Usage", in *Transactions on Large-Scale Data- and Knowledge-Centered Systems XXXVII*, ser. Lecture Notes in Computer Science, A. Hameurlain and R. Wagner, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2018, pp. 41–80, ISBN: 978-3-662-57932-9. DOI: 10.1007/978-3-662-57932-9_2. [Online]. Available: https://doi.org/10.1007/978-3-662-57932-9_2 (visited on 01/04/2019).

[135] A. Gerl and D. Pohl, "Critical Analysis of LPL according to Articles 12 - 14 of the GDPR", in *Proceedings of the 13th International Conference on Availability, Reliability and Security - ARES 2018*, Hamburg, Germany: ACM Press, 2018, pp. 1–9, ISBN: 978-1-4503-6448-5. DOI: 10/gfspqp. [Online]. Available: http://dl.acm.org/citation.cfm?doid=3230833.3233267 (visited on 01/04/2019).

[136] T. Lodge, A. Crabtree, and A. Brown, "Developing GDPR Compliant Apps for the Edge", in *Data Privacy Management, Cryptocurrencies and Blockchain Technology*, J. Garcia-Alfaro, J. Herrera-Joancomartí, G. Livraga, and R. Rios, Eds., vol. 11025, Cham: Springer International Publishing, 2018, pp. 313–328, ISBN: 978-3-030-00304-3 978-3-030-00305-0. DOI: 10.1007/978-3-030-00305-0_22. [Online]. Available: http://link.springer.com/10.1007/978-3-030-00305-0_22 (visited on 04/01/2019).

[137] D. Peras, "Guidelines for GDPR Compliant Consent and Data Management Model in ICT Businesses", in *29th International Conference of Central European Conference on Information and Intelligent Systems*, 2018, p. 9.

[138] J. Tom, E. Sing, and R. Matulevičius, "Conceptual Representation of the GDPR: Model and Application Directions", in *International Conference on Business Informatics Research*, ser. Lecture Notes in Business Information Processing, Springer, 2018, pp. 18–28, ISBN: 978-3-319-99951-7. DOI: 10/gft37v.

[139] P. Pullonen, J. Tom, R. Matulevičius, and A. Toots, "Privacy-enhanced BPMN: Enabling data privacy analysis in business processes models", *Software & Systems Modeling*, Jan. 30, 2019, ISSN: 1619-1366, 1619-1374. DOI: 10/gfv5x7. [Online]. Available: http://link.springer.com/10.1007/s10270-019-00718-z (visited on 02/03/2019).

[140] T. A. Coleti, M. Morandini, L. Vilela Leite Filgueiras, P. L. Pizzigatti Correa, I. G. de Oliveira, and C. R. S. C. de Barbosa, "Design Patterns to Support Personal Data Transparency Visualization in Mobile Applications", in *Human-Computer Interaction. Perspectives on Design*, M. Kurosu, Ed., vol. 11566, Cham: Springer International Publishing, 2019, pp. 46–62, ISBN: 978-3-030-22645-9 978-3-030-22646-6. DOI: 10.1007/978-3-030-22646-6_4. [Online]. Available: http://link.springer.com/10.1007/978-3-030-22646-6_4 (visited on 07/16/2019).

[141] M. Corrales, P. Jurčys, and G. Kousiouris, "Smart Contracts and Smart Disclosure: Coding a GDPR Compliance Framework", in *Legal Tech, Smart Contracts and Blockchain*, Springer, 2019, pp. 189–220.

[142] N. Havelange, M. Dumontier, B. Wouters, J. Linde, D. Townend, A. Riedl, and V. Urovi, "LUCE: A Blockchain Solution for monitoring data License accoUntability and ComplianceE", Aug. 6, 2019. arXiv: 1908.02287 [cs]. [Online]. Available: http://arxiv.org/abs/1908.02287 (visited on 08/21/2019).

[143] L. Sion, P. Dewitte, D. V. Landuyt, K. Wuyts, I. Emanuilov, P. Valcke, and W. Joosen, "An Architectural View for Data Protection by Design", in *2019 IEEE International Conference on Software Architecture (ICSA)*, Mar. 2019, pp. 11–20. DOI: 10/gf3czd.

[144] H. Gjermundrød, I. Dionysiou, and K. Costa, "privacyTracker: A Privacy-by-Design GDPR-Compliant Framework with Verifiable Data Traceability Controls", in *Current Trends in Web Engineering*, ser. Lecture Notes in Computer Science, Springer, Cham, Jun. 6, 2016, pp. 3–15, ISBN: 978-3-319-46962-1 978-3-319-46963-8. DOI: 10/gfxvs2. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-46963-8_1 (visited on 04/22/2018).

[145] D. Spagnuelo, C. Bartolini, and G. Lenzini, "Metrics for Transparency", in *Data Privacy Management and Security Assurance*, G. Livraga, V. Torra, A. Aldini, F. Martinelli, and N. Suri, Eds., vol. 9963, Cham: Springer International Publishing, 2016, pp. 3–18, ISBN: 978-3-319-47071-9 978-3-319-47072-6. DOI: 10.1007/978-3-319-47072-6_1. [Online]. Available: http://link.springer.com/10.1007/978-3-319-47072-6_1 (visited on 05/23/2019).

[146] V. Diamantopoulou, K. Angelopoulos, M. Pavlidis, and H. Mouratidis, "A Metamodel for GDPR-based Privacy Level Agreements", in *Proceedings of the ER Forum 2017 and the ER 2017 Demo Track Co-Located with the 36th International Conference on Conceptual Modelling (ER 2017), Valencia, Spain, - November 6-9, 2017.*, 2017, pp. 285–291. [Online]. Available: http://ceur-ws.org/Vol-1979/paper-08.pdf.

[147] M. Robol, M. Salnitri, and P. Giorgini, "Toward GDPR-Compliant Socio-Technical Systems: Modeling Language and Reasoning Framework", in *The Practice of Enterprise Modeling*, ser. Lecture Notes in Business Information Processing, Springer, Cham, Nov. 22, 2017, pp. 236–250, ISBN: 978-3-319-70240-7 978-3-319-70241-4. DOI: 10/gfxvs7. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-70241-4_16 (visited on 07/19/2018).

[148] F. Dalpiaz, E. Paja, and P. Giorgini, *Security Requirements Engineering: Designing Secure Socio-Technical Systems*. MIT Press, 2016.

[149] D. Basin, S. Debois, and T. Hildebrandt, "On Purpose and by Necessity: Compliance under the GDPR", in *Proceedings of Financial Cryptography and Data Security 2018*, Mar. 2018, p. 18.

[150] A. Palm, "Modelling Data Protection Vulnerabilities of Cloud Systems using Risk Patterns", RestAssured, Technical Report, 2018, p. 62.

[151] S. Schoenen, Z. Á. Mann, and A. Metzger, "Using Risk Patterns to Identify Violations of Data Protection Policies in Cloud Systems", in *Service-Oriented Computing – ICSOC 2017 Workshops*, L. Braubach, J. M. Murillo, N. Kaviani, M. Lama, L. Burgueño, N. Moha, and M. Oriol, Eds., vol. 10797, Cham: Springer International Publishing, 2018, pp. 296–307, ISBN: 978-3-319-91763-4 978-3-319-91764-1. DOI: 10.1007/978-3-319-91764-1_24. [Online]. Available: http://link.springer.com/10.1007/978-3-319-91764-1_24 (visited on 08/27/2019).

[152] *D5.1 Concept for End-User Privacy Policy Violation Detection*, 2018. [Online]. Available: https://restassuredh2020.eu/wp-content/uploads/2018/07/D5.1.pdf (visited on 08/27/2019).

[153] *D6.1 Methodology for Decentralized Data lifecycle Management*, 2018. [Online]. Available: https://restassuredh2020.eu/wp-content/uploads/2018/07/D6.1.pdf (visited on 08/27/2019).

[154] N. Gol Mohammadi, J. Leicht, N. Ulfat-Bunyadi, and M. Heisel, "Privacy Policy Specification Framework for Addressing End-Users' Privacy Requirements", in *Trust, Privacy and Security in Digital Business*, S. Gritzalis, E. R. Weippl, S. K. Katsikas, G. Anderst-Kotsis, A. M. Tjoa, and I. Khalil, Eds., vol. 11711, Cham: Springer International Publishing, 2019, pp. 46–62, ISBN: 978-3-030-27812-0 978-3-030-27813-7. DOI:

10.1007/978-3-030-27813-7_4. [Online]. Available: http://link.springer.com/10.1007/978-3-030-27813-7_4 (visited on 10/09/2019).

[155] *D7.1 RestAssured Security and Privacy Engineering Methodology*, 2018. [Online]. Available: https://restassuredh2020.eu/wp-content/uploads/2018/07/D7.1.pdf (visited on 08/27/2019).

[156] *D3.1 Guidelines on legal aspects*, 2016. [Online]. Available: http://www.operando.eu/upload/operando/moduli/D3.1-Guidelinesonlegalaspectsv2.0_77_289.pdf (visited on 10/09/2019).

[157] *D6.4 Final (Product) version of privacy enhanced tools*, 2017. [Online]. Available: http://www.operando.eu/upload/operando/moduli/D6.4FinalProductversionofprivacyenhancedtoolsv1.0_77_366.pdf (visited on 10/09/2019).

[158] *D6.7 Final (Product) Version of Security Aware Tools*, 2017. [Online]. Available: http://www.operando.eu/upload/operando/moduli/D6.7FinalProductVersionofSecurityAwareToolsv1.0_77_378.pdf (visited on 10/09/2019).

[159] *D1.1 Initial List of Main Requirements*, 2017. [Online]. Available: http://www.myhealthmydata.eu/wp-content/themes/Parallax-One/deliverables/D1.1_Initial-List-of-Main-Requirements.pdf (visited on 08/27/2019).

[160] D. Teodoro, E. Pasche, R. Mayer, and P. Ruch, *D4.2 Ontological Resources*, Apr. 30, 2018. [Online]. Available: http://www.myhealthmydata.eu/wp-content/themes/Parallax-One/deliverables/D4.2-MHMD-Ontological-Resources.pdf (visited on 08/27/2019).

[161] A. Bayle, M. Koscina, D. Manset, and O. Perez-Kempner, "When Blockchain Meets the Right to Be Forgotten: Technology versus Law in the Healthcare Industry", in *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, Santiago: IEEE, Dec. 2018, pp. 788–792, ISBN: 978-1-5386-7325-6. DOI: 10/gf9jn7. [Online]. Available: https://ieeexplore.ieee.org/document/8609693/ (visited on 10/08/2019).

[162] N. Sadeh, A. Acquisti, T. D. Breaux, L. F. Cranor, A. M. McDonald, J. R. Reidenberg, N. A. Smith, F. Liu, N. C. Russell, F. Schaub, *et al.*, "The usable privacy policy project", Technical Report, CMU-ISR-13-119, Carnegie Mellon University, 2013. [Online]. Available: http://ra.adm.cs.cmu.edu/anon/usr0/ftp/home/anon/isr2013/CMU-ISR-13-119.pdf (visited on 02/13/2017).

[163] A. Oltramari, D. Piraviperumal, F. Schaub, S. Wilson, S. Cherivirala, T. B. Norton, N. C. Russell, P. Story, J. Reidenberg, and N. Sadeh, "PrivOnto: A semantic framework for the analysis of privacy policies", *Semantic Web*, vol. 9, no. 2, M. d'Aquin, S. Kirrane, S. Villata, M. d'Aquin, S. Kirrane, and S. Villata, Eds., pp. 185–203, Jan. 24, 2018, ISSN: 22104968, 15700844. DOI: 10/gdfqnk. [Online]. Available: http://www.medra.org/servlet/aliasResolver?alias=iospress&doi=10.3233/SW-170283 (visited on 04/15/2018).

[164] H. Harkous, K. Fawaz, R. Lebret, F. Schaub, K. G. Shin, and K. Aberer, "Polisis: Automated analysis and presentation of privacy policies using deep learning", in *27th USENIX Security Symposium (USENIX Security 18)*, 2018, pp. 531–548.

[165] T. Linden, R. Khandelwal, H. Harkous, and K. Fawaz, "The Privacy Policy Landscape After the GDPR", Sep. 22, 2018. arXiv: 1809.08396 [cs]. [Online]. Available: http://arxiv.org/abs/1809.08396 (visited on 05/31/2019).

[166] M. Galle, A. Christofi, and H. Elsahar, "The Case for a GDPR-specific Annotated Dataset of Privacy Policies", in *Proceedings of the PAL: Privacy-Enhancing Artificial Intelligence and Language Technologies As Part of the AAAI Spring Symposium Series (AAAI-SSS 2019)*, 2019, p. 3. [Online]. Available: http://ceur-ws.org/Vol-2335/1st_PAL_paper_5.pdf.

[167] S. Wilson, F. Schaub, A. A. Dara, F. Liu, S. Cherivirala, P. Giovanni Leon, M. Schaarup Andersen, S. Zimmeck, K. M. Sathyendra, N. C. Russell, T. B. Norton, E. Hovy, J. Reidenberg, and N. Sadeh, "The Creation and Analysis of a Website Privacy Policy Corpus", in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 1330–1340. DOI: 10/gf9t98.

[168] W. B. Tesfay, P. Hofmann, T. Nakamura, S. Kiyomoto, and J. Serna, "PrivacyGuide: Towards an Implementation of the EU GDPR on Internet Privacy Policy Evaluation", in *Proceedings of the Fourth ACM International Workshop on Security and Privacy Analytics*, ser. IWSPA '18, New York, NY, USA: ACM, 2018, pp. 15–21, ISBN: 978-1-4503-5634-3. DOI: 10/gfxvsx. [Online]. Available: http://doi.acm.org/10.1145/3180445.3180447 (visited on 04/15/2018).

[169] M. A. Grando, A. Boxwala, R. Schwab, and N. Alipanah, "Ontological Approach for the Management of Informed Consent Permissions", in *2nd International Conference on Healthcare Informatics, Imaging and Systems Biology*, IEEE, Sep. 2012, pp. 51–60, ISBN: 978-0-7695-4921-7 978-1-4673-4803-4. DOI: 10/gfxvsr. [Online]. Available: http://ieeexplore.ieee.org/document/6366189/ (visited on 02/13/2017).

[170] M. Lizar and D. Turner, "Consent Receipt Specification v1.1.0", Kantara Initiative, 2017, p. 29. [Online]. Available: https://docs.kantarainitiative.org/cis/consent-receipt-specification-v1-1-0.pdf.

[171] *D2.1 Multistakeholder Specifications*, 2019.

[172] 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 *Trust, Privacy and Security in Digital Business*, S. Gritzalis, E. R. Weippl, S. K. Katsikas, G. Anderst-Kotsis, A. M. Tjoa, and I. Khalil, Eds., ser. Lecture Notes in Computer Science, Springer International Publishing, 2019, pp. 78–93, ISBN: 978-3-030-27813-7.

[173] G. Spindler, A. Z. Horváth, and L. Dalby, *D3.1 General Legal Aspects*, 2017.

[174] D. Roio and M. Sachy, *D3.5 Initial definition of Smart Rules and Taxonomy*, 2018.

[175] *D1.8 Legal frameworks for digital commons DECODE OS and legal guidelines*, 2017.

[176] *D2.1 - Use cases analysis and user scenarios*, 2018. [Online]. Available: https://www.poseidon-h2020.eu/wp-content/uploads/2019/08/PoSeID-on_D2.1-Use-cases-analysis-and-user-scenarios-v1.00.pdf (visited on 10/09/2019).

[177] *D3.1 PoSeID-on blockchain - Interim implementation*, 2019. [Online]. Available: https://www.poseidon-h2020.eu/wp-content/uploads/2019/08/D3.1_final-version_POSEIDON_v10.pdf (visited on 10/09/2019).

[178] *D4.3 Risk Management Module & Personal Data Analyser - Interim implementation*, 2019. [Online]. Available: https://www.poseidon-h2020.eu/wp-content/uploads/2019/08/D4.3-RMM-and-PDA-V1.0-Final.pdf (visited on 10/09/2019).

[179] A. Gangemi, S. Peroni, D. Shotton, and F. Vitali, "The Publishing Workflow Ontology (PWO)", *Semantic Web*, vol. 8, no. 5, V. de Boer and A. Ławrynowicz, Eds., pp. 703–718, Apr. 6, 2017, ISSN: 22104968, 15700844. DOI: 10/ggh67t. [Online]. Available: https://www.medra.org/servlet/aliasResolver?alias=iospress&doi=10.3233/SW-160230 (visited on 01/20/2020).

[180] S. M. Gurk, C. Abela, and J. Debattista, "Towards Ontology Quality Assessment", *Joint proceedings of the MEPDaW*, p. 12, 2017.

[181] D. Vrandecic, "Ontology evaluation", PhD Thesis, Karlsruhe Institute of Technology, 2010. [Online]. Available: http://digbib.ubka.uni-karlsruhe.de/volltexte/1000018419.

[182] M. Hintze and G. LaFever, *Meeting Upcoming GDPR Requirements While Maximizing the Full Value of Data Analytics*, 2017. [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2927540 (visited on 05/24/2017).

[183] C. Debruyne, J. Riggio, O. De Troyey, and D. O'Sullivan, "An Ontology for Representing and Annotating Data Flows to Facilitate Compliance Verification", in *2019 13th International Conference on Research Challenges in Information Science (RCIS)*, May 2019, pp. 1–6. DOI: 10/ggkj8n.

[184] S. Cox and C. Little, "Time ontology in OWL", *World Wide Web Consortium. Retrieved from https://www. w3. org/TR/owl-time*, 2017.

[185] R. Iannella and S. Villata. (Feb. 15, 2018). ODRL Information Model 2.2, [Online]. Available: https://www.w3.org/TR/odrl-model/ (visited on 09/19/2018).

[186] P. Bonatti, S. Kirrane, R. Wenning, A. Corazza, C. Galdi, A. Apicella, W. Dullaert, P. Raschke, and L. Sauro, *D1.7 Policy, transparency and compliance guidelines V2*, Jul. 31, 2018. [Online]. Available: https://www.specialprivacy.eu/images/documents/SPECIAL_D1.7_M17_V1.0.pdf (visited on 10/12/2018).

[187] P. A. Bonatti, S. Kirrane, I. M. Petrova, L. Sauro, and E. Schlehahn, *The SPECIAL Vocabularies v0.1*, 2018. [Online]. Available: https://www.specialprivacy.eu/vocabs.

[188] V. Leone, L. D. Caro, and S. Villata, "Legal Ontologies and How to Choose Them: The InvestigatiOnt Tool", in *Proceedings of the ISWC 2018 Posters & Demonstrations, Industry and Blue Sky Ideas Tracks Co-Located with 17th International Semantic Web Conference (ISWC 2018), Monterey, USA, October 8th - to - 12th, 2018.*, 2018. [Online]. Available: http://ceur-ws.org/Vol-2180/paper-36.pdf.