UNIVERSIDADE FEDERAL FLUMINENSE

DANIEL PRETT CAMPAGNA

A DATA PROVENANCE MODEL FOR THE GENERAL DATA PROTECTION REGULATION

NITERÓI

2026

UNIVERSIDADE FEDERAL FLUMINENSE

DANIEL PRETT CAMPAGNA

A DATA PROVENANCE MODEL FOR THE GENERAL DATA PROTECTION REGULATION

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Computação da Universidade Federal Fluminense como requisito parcial para a obtenção do Grau de Mestre em Computação. Área de concentração: Ciência da Computação

Orientador:
Vanessa Braganholo

Co-orientador:
Altigran Soares da Silva

NITERÓI

2026

Ficha catalográfica automática - SDC/BEE
Gerada com informações fornecidas pelo autor

C186d Campagna, Daniel Prett
A data provenance model for the General Data Protection Regulation / Daniel Prett Campagna. - 2026.
143 f.

Orientador: Vanessa Braganholo.
Coorientador: Altigran Soares da Silva.
Dissertação (mestrado)-Universidade Federal Fluminense, Instituto de Computação, Niterói, 2026.

1. Direito à privacidade. 2. Banco de dados. 3. Ontologia (Computação). 4. Produção intelectual. I. Braganholo, Vanessa, orientadora. II. Silva, Altigran Soares da, coorientador. III. Universidade Federal Fluminense. Instituto de Computação. IV. Título.

CDD - XXX

Bibliotecário responsável: Debora do Nascimento - CRB7/6368

Daniel Prett Campagna

A DATA PROVENANCE MODEL FOR THE GENERAL DATA PROTECTION REGULATION

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Computação da Universidade Federal Fluminense como requisito parcial para a obtenção do Grau de Mestre em Computação. Área de concentração: Ciência da Computação

Aprovada em Junho de 2026.

BANCA EXAMINADORA

Profa. Vanessa Braganholo Murta - Orientadora, UFF

Prof. Altigran Soares da Silva - Coorientador, UFAM

Profa. Flavia Cristina Bernardini, UFF

Prof. Javam de Castro Machado, UFC

Niterói

2026

Acknowledgments

No man is an island, the poet mused. Likewise, no research is an island entirely itself. Every researcher owes a debt to many others who contributed to shaping their views and who laid the groundwork for their discoveries.

First and foremost, I thank my advisor, Prof. Vanessa Braganholo, for the guidance, patience, and encouragement throughout this work, and my co-advisor, Prof. Altigran Soares da Silva, for the careful reading and the many insightful suggestions that sharpened both the research and its presentation.

I am also grateful to the Universidade Federal Fluminense and to the Programa de Pós-Graduação em Computação for the environment and the resources that made this dissertation possible, as well as to the professors and colleagues whose discussions, directly or indirectly, contributed to this result.

Finally, I thank my family and friends for their unconditional support during this journey.

Resumo

A aprovação do Regulamento Geral de Proteção de Dados (GDPR) trouxe uma revolução na forma como tratamos os dados pessoais. O GDPR aumenta a participação dos indivíduos no tratamento de seus dados e também introduz desafios técnicos, cujo descumprimento pode acarretar em multa de até 4% do faturamento anual da empresa descumpridora. Para facilitar os esforços de conformidade, as autoridades supervisoras publicaram documentos de orientação que traduzem as disposições da lei em listas de questões de conformidade. Técnicas de proveniência de dados têm sido usadas com sucesso para representar e verificar essas questões; ainda assim, uma parcela substancial delas permanece sem representação e, portanto, não pode ser verificada automaticamente. Para reduzir essa lacuna, nesta dissertação, estendemos a GDPRov, a ontologia de proveniência proposta por Pandit, para abordar quatro questões de conformidade ainda não cobertas do GDPR Guidance for SMEs, publicado pela Data Protection Commission da Irlanda: o período de retenção de cada categoria de dados pessoais, os procedimentos para manter os dados pessoais exatos e atualizados sem demora, as políticas de retenção que garantem que os dados não sejam mantidos por mais tempo do que o necessário e a eliminação ou anonimização sistemática dos dados pessoais quando não são mais legalmente exigidos. Reutilizando construtos do Data Privacy Vocabulary, da ontologia Time e da Semantic Sensor Network ontology, introduzimos uma representação causal que vincula o cumprimento da finalidade de coleta de dados ao acionamento da eliminação ou anonimização, uma relação que nenhum modelo de proveniência para o GDPR representava. Validamos o modelo estendido em três cenários reais de uso sobre a base de dados do programa de pós-graduação SAPOS, sob a política de privacidade do Trinity College Dublin, respondendo a cada questão com consultas SPARQL independentes do caso de uso. Como resultado, a cobertura do GDPR Guidance for SMEs aumenta de 32 para 36 de suas 65 questões, um aumento de 12,5% em relação ao modelo anterior, e uma análise das 29 questões ainda não cobertas revela cinco categorias reutilizáveis que apontam caminhos sistemáticos para extensões futuras.

Palavras-chave: GDPR, proveniência, conformidade.

Abstract

The approval of the General Data Protection Regulation (GDPR) brought a revolution in the way we treat personal data. The GDPR increases individuals' participation in the treatment of their data, and it also introduces technical challenges, whose failure can lead to a fine of up to 4% of the annual revenue of the offending enterprise. To facilitate compliance efforts, supervisory authorities have published guidance documents that translate the law provisions into checklists of compliance questions. Data provenance techniques have been successfully used to represent and check these questions, yet a substantial portion of them remains unrepresented and thus cannot be verified automatically. To narrow this gap, in this dissertation, we extend GDPRov, the provenance ontology proposed by Pandit, to address four previously uncovered compliance questions from the GDPR Guidance for SMEs, published by the Data Protection Commission of Ireland: the retention period of each personal-data category, procedures to keep personal data accurate and up to date without delay, retention policies that ensure data are held no longer than necessary, and the systematic erasure or anonymization of personal data once they are no longer legally required. Reusing constructs from the Data Privacy Vocabulary, the Time ontology, and the Semantic Sensor Network ontology, we introduce a causal representation that links the fulfilment of a data-collection purpose to the triggering of deletion or anonymization, a relationship no prior GDPR provenance model represents. We validate the extended model over three real use-case scenarios on the SAPOS graduate-program database under the Trinity College Dublin privacy policy, answering each question with use-case-independent SPARQL queries. As a result, the coverage of the GDPR Guidance for SMEs rises from 32 to 36 of its 65 questions, a 12.5% increase over the prior model, and an analysis of the remaining 29 uncovered questions reveals five reusable categories that chart systematic pathways for future extensions.

Keywords: General Data Protection Regulation, Data provenance, conformance.

List of Figures

List of Tables

List of Abbreviations and Acronyms

AEPD	:	Agencia Española de Protección de Datos;
ANPD	:	Autoridade Nacional de Proteção de Dados;
BDSG	:	Bundesdatenschutzgesetz;
BERT	:	Bidirectional Encoder Representations from Transformers;
BFO	:	Basic Formal Ontology;
BPMN	:	Business Process Model and Notation;
CACM	:	Common Assurance and Certification Metamodel;
CCPA	:	California Consumer Privacy Act;
CDMM	:	Consent and Data Management Model;
CNIL	:	Commission Nationale de l'Informatique et des Libertés;
CQ	:	Competency Question;
CSM	:	Common Semantic Model;
CSM-ROPA	:	Common Semantic Model for ROPA;
DCAT	:	Data Catalog Vocabulary;
DFD	:	Data Flow Diagram;
DGA	:	Data Governance Act;
DPA	:	Data Protection Authority;
DPC	:	Data Protection Commission;
DPCat	:	Data Processing Catalog;
DPIA	:	Data Protection Impact Assessment;
DPO	:	Data Protection Officer;
DPV	:	Data Privacy Vocabulary;
DPVCG	:	Data Privacy Vocabularies and Controls Community Group;
DSL	:	Domain Specific Language;
EC	:	European Commission;
ECHR	:	European Convention on Human Rights;
ECLI	:	European Case Law Identifier;

EDPB : European Data Protection Board;
EEA : European Economic Area;
ELI : European Legislation Identifier;
EU : European Union;
GDPR : General Data Protection Regulation;
GPT : Generative Pre-trained Transformer;
HIPAA : Health Insurance Portability and Accountability Act;
ICO : Information Commissioner's Office;
IEC : International Electrotechnical Commission;
IEEE : Institute of Electrical and Electronics Engineers;
IoT : Internet of Things;
ISO : International Organization for Standardization;
JSON : JavaScript Object Notation;
JSON-LD : JSON for Linking Data;
KG : Knowledge Graph;
LGPD : Lei Geral de Proteção de Dados;
LKIF : Legal Knowledge Interchange Format;
LLM : Large Language Model;
NIST : National Institute of Standards and Technology;
NISTIR : NIST Internal/Interagency Report;
OCL : Object Constraint Language;
ODRL : Open Digital Rights Language;
OOPS : OntOlogy Pitfall Scanner;
OPPO : Ontology for Privacy Policies of OSNs;
OSN : Online Social Network;
OWL : Web Ontology Language;
P3P : Platform for Privacy Preferences;
PaCW : Privacy-aware Clinical Workflow;
PbD : Privacy by Design;
PLR : Policy Logic Reasoner;
PME : Petite et Moyenne Entreprise / Pequena e Média Empresa;
P-Plan : Plan Ontology;
PRIAM : PRIvacy Assessment Model;

PROV : Provenance (W3C family);
PROV-DM : PROV Data Model;
PROV-O : PROV Ontology;
QCA : Qualitative Content Analysis;
RDF : Resource Description Framework;
RDFS : RDF Schema;
RGPD : Règlement Général sur la Protection des Données / Reglamento General de Protec
RO : Research Objective;
ROPA : Register of Processing Activities;
RQ : Research Question;
SAA : Student Administration and Affairs;
SAPOS : Sistema de Acompanhamento de Pós-Graduação;
SARs : Subject Access Requests;
SCHAL : Shapes Constraint Language (variant);
SHACL : Shapes Constraint Language;
SKOS : Simple Knowledge Organization System;
SME : Small and Medium Enterprise;
SOSA : Sensor, Observation, Sample, and Actuator;
SoT : Source of Truth;
SPARQL : SPARQL Protocol and RDF Query Language;
SPECIAL : Scalable Policy-aware Linked Data Architecture For Privacy, Transparency and Con
SQL : Structured Query Language;
SSN : Semantic Sensor Network;
SSoT : Single Source of Truth;
STS-ml : Socio-Technical Security modeling language;
SWRL : Semantic Web Rule Language;
TCD : Trinity College Dublin;
TCU : Trinity College University ontology;
TPE : Très Petite Entreprise;
UFF : Universidade Federal Fluminense;
UK : United Kingdom;
UML : Unified Modeling Language;
URI : Uniform Resource Identifier;
US : United States;

W3C : World Wide Web Consortium;

WP29 : Article 29 Working Party;

XML : eXtensible Markup Language;

XSD : XML Schema Definition;

Contents

Chapter 1

Introduction

1.1 Motivation and Problem

In the last decade, European lawmakers have updated old privacy regulations. Rather than conceiving privacy in terms of bureaucratic requirements, they have shifted towards the individual's conception of what privacy means, as Christopher Kuner, professor of Law at the Vrije Universiteit Brussel, has argued (67). This movement gives birth to the General Data Protection Regulation (GDPR) (1). By embodying that premise, the GDPR lays down rules to ensure the individual's decision is at the center regarding the processing of their data (personal data (1, Art. 4 (1))). Under GDPR, data about a person is no longer the property of those who possess the data (controller or processor (1, Art. 4 (7), 4 (8))); actually, such data belong to the individual whose such data can identify (data subject (1, Art. 4 (1))). Consequently, every processing involving personal data often needs the data subject's consent to run (one of the legal basis of GDPR (1, Art. 6 (1))). Additionally, such processing can also be restricted may the data subject object to it (Right to Object (1, Arts. 12-23)). To summarize, the data subject has [almost] complete control over their data.

In addition to this reconceptualization of the meaning of privacy (gravitating around the data subject's will), this movement is as impactful as previous revolutions in scope; it goes as far as Europeans can go: reaching every company that provide services for an European citizen, being painful for all those who [stubbornly] do not comply with it. Indeed, the GDPR lays down severe fines for non-compliant bodies (i.e., up to 20 million euros or 4% of the total worldwide annual revenue, whichever is higher (1, Art. 83)). The law also scopes those bodies not established in the European Union (EU) jurisdiction which have into their activities the processing of European data (1, Art. 3).

Beyond the territorial scopes of the GDPR, its revolutionary spirit has also inspired 62%

of the countries towards promulgating their own privacy law. Since April 2021, 177 countries have enacted data privacy laws to protect their residents' data (57; 56; 55; 58). Despite their particularities, these laws were influenced by the GDPR premise (27). For example, Brazil's General Data Protection Law (LGPD) (24) presents broad similarities with GDPR inter alias the definition of personal data, some legal basis for processing, and some individuals' rights[1] (5; 45; 73).

Thus far, we have a revolutionary and severe law regarding implementation complexity and fines. Despite being enacted in the European Union territories, it has also affected many companies worldwide, either by its rules or by influencing third-countries legislations. But how have companies dealt with these requirements? Are there still open issues that motivate some investigation?

Although the GDPR is celebrating eight years, open issues remain. The number of fines imposed yearly increases each period (2). A 2019 survey has reported that two-thirds of the small and medium enterprises (SMEs) in Europe do not completely describe their data processing activities in clear and plain language to data subjects (46), such as Articles 12 and 13 obligate (1). A 2019 survey has identified as the foremost priority for organizations the preparation of internal procedures to meet the principle of transparency and accountability (70). This principle has a straightforward relationship with the data subject's right to access their data (1, Art. 15), which empowers the data subject to obtain from the controller confirmation of whether or not personal data concerning them are being processed and, where that is the case, the information. Indeed, this right is responsible for the major complaints received from individuals, according to Annual Reports published in 2021, 2022, 2023, and 2024 by the Data Protection Commission (DPC), Ireland's independent supervisory authority (35). This right burdens the controllers with the responsibility to answer compliance questions, such as, what are the processes involved with this data? What are the purposes of these processes? What are the categories of personal data involved? What are the time limits for the erasure of these data? And so on.

A branch of the literature addressing these questions has been proposing efforts to collect and manage the records of processing activities. Indeed, Article 30 reinforces these kinds of efforts (1), which lie in the field of interest of research in data provenance. Data provenance can be understood as historical information about processes, persons, and data involved in an activity (3). Data provenance has a large tradition of publications promoting its usage in scientific contexts due to the benefits of tracking the activities involved in experi-

[1] In Brazil, sparks of the right of data portability go back to 2007, with Resolution 460/07, also known as the General Portability Regulation of Anatel, that ensures such right for telephone numbers.

ments (99), facilitating the reproducibility and the reasoning about the outcomes obtained from experiments (49; 29; 32). Due to these benefits, it is possible to encounter a considerable number of approaches in literature endorsing the use of provenance to capture the records of processing regarding the GDPR (90; 85; 112; 16; 48) so that controllers can answer those questions.

Part of the literature about data provenance addresses the GDPR concerns by representing these records to answer compliance questions about the principle of transparency and accountability. Depending on the knowledge-representation model proposed, some questions can either be or not be answered. For instance, Ujcich et al. (112) have proposed a data provenance model to represent the data processing and the GDPR concepts by extending the PROV (3) data model into relevant GDPR concepts such as legal basis, controller, purpose, personal data, etc. Ujcich et al. show that their model can answer whether or not a piece of data was used after the data subject withdraws her consent; and, where that is the case, who used it and which portion of such data was used. It also addresses processor-perspective questions. Indeed, a recent revision of the literature (85) have appointed this model as one of the complete models for representing activities, legal basis, rights, purposes, processing, data sharing, etc. However, this same revision has also appointed it is not clear how this model would answer complex questions and also that this model fails to represent ex-ante information, which is mandatory in some cases [2].

Another example came from Pandit publications (96; 91; 92; 93; 89; 95; 94; 87; 85). His contributions make significant efforts in demonstrating the benefits of using provenance and semantic web technologies (85; 90; 96). Among the representations he has proposed, he presents GDPRov as a GDPR provenance ontology, which aims to represent the provenance of consent and activities associated with processing personal data. GDPRov advances the state of the art since it is capable of representing ex-ante information of activities. Hence, within GDPRov, before the process execution, the controller can answer questions related to Data Protection Impact Assessment (1, Art. 35). Pandit evaluates GDPRov by mapping compliance questions extracted from official guideline documents (36) into machine-readable questions and constraints (i.e., utilizing SPARQL, SCHAL) and demonstrates that such representation makes it feasible to answer some of those questions. Conversely, despite the exhaustive Pandit's contribution, GDPRov could answer only half of these questions. He ar-

[2] Pandit argues that representation of "ex-ante information about processes is essential in the evaluation of compliance", in particular, to meet the Data Protection Impact Assessment (DPIA) (1, Art. 35). DPIA specifies particular cases where the compliance assessment of processing should be carried out before this processing take place. In provenance, we have identified the term *prospective provenance* to identify the ex-ante information. More precisely, prospective provenance specifies a process's procedure calls and data dependencies. You can see more about this definition in (33)

gues that less than the other half could not be represented by provenance.

The motivation of this work lies in the importance of the GDPR, the still large numbers of fines imposed yearly due to the lack of demonstrating compliance (in particular those cases related to the principle of transparency and accountability), the well-established literature in assisting the compliance demonstration (through the representation of the information utilized in answering the compliance question), and the identified opportunities in these prior work by not covering all of these questions.

Throughout this dissertation, we refer to these compliance questions as competency questions (CQ), the term adopted by the ontology-engineering literature to denote the questions that an ontology must be able to answer (78). Our contribution's validity rests specifically on four of those uncovered questions: CQ8 (the retention period of each personal-data category), CQ28 (procedures to keep personal data accurate and up to date without delay), CQ29 (retention policies that ensure data are held no longer than necessary), and CQ51 (the systematic erasure or anonymization of personal data once they are no longer legally required). We extract and analyze the terms present in each question and propose new concepts for GDPRov so that the model can address them. Among these additions, we introduce a causal representation that links the fulfilment of a data-collection purpose to the triggering of deletion or anonymization, a relationship that no prior GDPR provenance model represents. Following traditional ontology-engineering methodologies to extend conceptual models, we evaluate our contribution by instantiating three use-case scenarios in a real database supported by a real privacy policy. Beyond the four questions, we also chart the remaining uncovered ones into five reusable categories and deliver the extension as a publicly available model with use-case-independent SPARQL query patterns, corroborating, within this narrower scope, the thesis that data provenance models can be extended to expand the corpus of official compliance questions answered by such techniques.

1.2 Research Objective

1.2.1 Research Question

The literature has employed data provenance to demonstrate compliance with the GDPR, representing, querying, and validating the information utilized to answer compliance questions. This coverage, however, remains partial: a substantial portion of the official compliance questions is still not represented by the current models. In light of the persistent demand for compliance automation introduced beforehand, we inquire:

Research Question (RQ): To what extent can data provenance models answer the official GDPR compliance questions that current models leave uncovered?

Premise. Before any efforts toward seeking the answer of **RQ**, it is necessary to highlight the premise behind this question in light of the motivation introduced beforehand. This research question is tied on the premise that the extension of the current data provenance models to represent uncovered obligations might contribute to the reduction of the large numbers of fines imposed yearly due to the lack of demonstrating compliance (in particular, those fines related to the principle of transparency and accountability)[3].

1.2.2 Scope

Translating the GDPR's basic principles into software requirements remains challenging, error-prone, and a non-trivial task (81; 54). This is because legal terms like *privacy by design* (1, Art. 25) are often too abstract and their encoding usually demands some personal interpretation and inferences that may lead to inconsistencies (15; 66; 74)[4]. For the sake of the controllers, supervisory authorities, such as the Irish Data Protection Commission (DPC), have provided guidance documents and checklists that attenuate these translation issues by offering more objective and practical instructions for controllers to achieve and assess their own compliance level.

Among such guidance documents, this dissertation adopts the obligations elicited by the GDPR Guidance for SMEs (36), published by the DPC as Chapter 6 of the GDPR establishes. The purposes of a supervisory authority include inter alias conducting investigations and law enforcement, where necessary, and publishing high-quality guidance to engage and raise awareness of such legislators. Hence, it seems reasonable to adopt this document to identify the obligations since the DPC is the one that inquires and enforces the law.

[3]The authors are conscious of the impossibility of demonstrating this premise is met in this work. However, it is an inevitable premise. We start our argumentation from a context where many fines are being imposed due to the lack of demonstration of compliance. Since that is the problem we place, the expectation is that this work addresses this problem. The authors choose a position of transparency and honesty by opting to highlight the vulnerabilities and limitations of this work, despite the similarity of the reasons behind part of these fines and contributions presented here: i.e., demonstrating compliance.

[4](15) shed light on a linguistic issue: lawyers, makers, and courts endeavor to describe things using legal terminology. In that process, they propose and enrich a common legal vocabulary capable of describing such things. At the same time, software engineers and IT stakeholders have also evolved their language to describe the things they are interested in. This latter vocabulary is composed of different symbols and grammars; it goes beyond common human languages (e.g., English, Portuguese), encompassing formal languages and rules that establish the relation between the symbols. This linguistic problem concerns the translation of terms across different realms of knowledge, the impossibility of an existing final methodology that ensures identical translation of terms, and the difficulties in measuring, comparing, and evaluating different translations.

GDPR Guidance for SMEs is a guideline composed of checklists of questions designed to assist the small and medium enterprise (SMEs) sector in getting data-protection compliance (36). This document was published twice. The 2017 and 2019 versions list 63 and 65 obligations, respectively. These obligations are presented in a format of questions grouped into eight categories related to the law, namely: general, personal data, data subjects' rights, questions about accuracy and retention time of data, transparency requirements, data breaches and security, and international transfers.

To the best of our knowledge, GDPRov is the only model to utilize this document for its evaluation (85). The author demonstrates how those obligations can be represented as a data provenance model constructed over Semantic Web technologies to assist the assessment of achieving compliance. Aligning our analysis with the 2019 version of the Guidance, which lists 65 obligations, GDPRov addresses 32 of them, leaving 33 uncovered; 13 of those 33, Pandit has argued, have no relation to provenance.

As far as we know, no other contribution has utilized this document to identify the obligations. Under the 2019 reference frame, this leaves 20 of the 33 uncovered obligations of the GDPR Guidance for SMEs (36) (those Pandit did not classify as outside the scope of provenance) still open to be represented through provenance-based approaches.

Of these uncovered obligations, this work addresses four chossen for two reinforcing reasons rather than as an arbitrary sample of the uncovered set. First, together they span the full lifecycle of stored personal data: how long each category may be retained (CQ8), keeping it accurate while retained (CQ28), the policy that bounds its retention (CQ29), and its erasure or anonymization once it is no longer needed (CQ51). Treating them as a single unit rather than in isolation lets the same constructs be shared across questions — CQ51, for instance, is answered by reusing the constructs introduced for CQ8 under an inverse query, without adding any new element. Second, this cluster contains the single hardest gap the field has left open: representing the causal link between the fulfilment of a data-collection purpose and the triggering of deletion. Pandit identified this relationship but explicitly scoped it out of his contribution because of its difficulty, and no provenance model reviewed in Chapter 3 represents it. Selecting these questions therefore targets a foundational, unclaimed problem rather than incremental coverage. This advances the total number of covered competency questions (CQ) from 32 to 36. Specifically, we address the following:

- **CQ8**: For each category of personal data, list the period for which the data will be retained (e.g. one month? one year?). As a general rule, data must be retained for no longer than is necessary for the purpose for which it was collected in the first place.

- **CQ28**: Are procedures in place to ensure personal data is kept up to date and accurate, and where a correction is required, the necessary changes are made without delay?
- **CQ29**: Are retention policies and procedures in place to ensure data are held for no longer than is necessary for the purposes for which they were collected?
- **CQ51**: Are personal data systematically destroyed, erased, or anonymised when they are no longer legally required to be retained?

The four competency questions addressed here are not idiosyncratic to the Irish DPC checklist (36). Each one operationalizes a core GDPR obligation: CQ8 and CQ29 instantiate the storage-limitation principle (1, Art. 5(1)(e)), CQ28 the accuracy principle and the right to rectification (1, Art. 5(1)(d), Art. 16), and CQ51 the right to erasure together with storage limitation (1, Art. 17, Recital 26). Because they are principle-derived rather than checklist-specific, the same questions recur, often almost verbatim, across the official compliance instruments published by other supervisory authorities and privacy frameworks. Table 1.1 illustrates this convergence with verbatim items drawn from the Spanish AEPD compliance checklist (10), the U.S. NIST Privacy Framework (75), and the German BayLDA data-protection checklist for small enterprises (17). Equivalent obligations also appear in the UK ICO data-protection self-assessment (63), the French CNIL *check-list RGPD pour les TPE/PME* (34), and, under a non-GDPR regime, the Brazilian ANPD guidance for small-scale processing agents (14), consistent with the storage-limitation and data-quality provisions of the LGPD (24). This cross-instrument, cross-jurisdiction recurrence indicates that the term-decomposition we adopt is not bound to a single checklist: any principle-derived compliance question can be decomposed into the same kind of terms regardless of the instrument that phrases it.

Once we have scoped which obligations we intend to address and identified that the current literature does not fully cover these obligations, we can formulate the **RO** of proposing a new data provenance model capable of representing these uncovered obligations, and verify later, in Chapter 5, to what extent such a model answers the **RQ**.

1.2.3 Research Objective

Research Objective (RO): Propose an extension to an existing data provenance model that represents, queries, and validates the information required to answer a set of four previously uncovered compliance questions of the GDPR Guidance for SMEs (CQ8, CQ28, CQ29, and CQ51).

Table 1.1: Verbatim items from independent official compliance instruments in other jurisdictions and frameworks matching each of the four addressed competency questions (Spanish and German items in the original language).

CQ (Irish DPC (36))	AEPD (Listado de Cumplimiento, ES (10))	NIST Privacy Framework v1.0, US (75)	BayLDA Datenschutz-Checkliste, DE (17)
CQ8	"Se informa de los criterios utilizados para determinar el plazo de conservación"	"… conditions on data processing such as data uses or retention periods …"	"Festlegung Löschfristen"; "… acht Jahre bei Rechnungen"
CQ28	"Los datos personales se mantienen exactos"; "Se rectifican los datos personales inexactos sin dilación indebida"	"Data corrections or deletions can be communicated to individuals or organizations … in the data processing ecosystem."	"Betroffenenrechte wie … Berichtigungs- oder Löschrechte"
CQ29	"Se mantienen durante más tiempo del necesario respecto de la finalidad"	"Policies, processes, and procedures for enabling data review … and deletion are established"	"Die rein datenschutzrechtliche Speicherdauer ergibt sich aus der Erforderlichkeit zur Erreichung des Zwecks"
CQ51	"… se suprimirán o devolverán los datos personales una vez finalice la prestación …"	"Data are destroyed according to policy." (CT.DM-P)	"Löschung nicht mehr benötigter Daten, für die keine gesetzlichen Aufbewahrungspflichten"

Within this work, all efforts aimed at proposing a data provenance model that reutilizes concepts well-established in the literature so that the private stakeholders who are subject to the GDPR, such as the controllers, can feed their provenance-collector framework with a new vocabulary that makes it possible to demonstrate compliance with the GDPR according to the GDPR Guidance for SMEs document (36).

1.3 Research Methodology

The roadmap to achieve the objective of this study is based on a combination of well-established knowledge-engineering methodologies (78; 107), slightly adapted into a four-step guideline for our case. Since the detailed application of these methodologies belongs to the construction of the model itself, we present it in the Methodology section of Chapter 4 (Section 4.1), where each step is described along with its corresponding chapter.

1.4 Contributions

The main contribution of this dissertation is an extension to the GDPRov ontology that enables the automated assessment of four official GDPR compliance questions that previous

provenance models did not cover: CQ8 (the retention period of each personal-data category), CQ28 (procedures to keep personal data accurate and up to date without delay), CQ29 (retention policies that ensure data are held no longer than necessary), and CQ51 (the systematic erasure or anonymization of personal data once they are no longer legally required). To this end, we reuse constructs from the Data Privacy Vocabulary, the Time ontology, and the Semantic Sensor Network ontology, and we introduce a causal representation that links the fulfilment of a data-collection purpose to the triggering of deletion or anonymization, a relationship that no prior GDPR provenance model represents. We further validate the extended model on three real use-case scenarios over the SAPOS graduate-program database under the Trinity College Dublin privacy policy, with each target question answered by SPARQL queries that are independent of the use case. As a result, the coverage of the GDPR Guidance for SMEs rises from 32 to 36 of its 65 official questions, a 12.5% increase over the prior model. Beyond the four addressed questions, our analysis of the remaining 29 uncovered questions reveals five reusable categories that suggest systematic pathways for future extensions.

1.5 Organization

The rest of this dissertation is structured as follows:

Chapter 2 — Background: The GDPR Concepts and Requirements, and the Data Provenance Models. This chapter contains a summary of the terms and concepts utilized in the literature that are necessary to understand this study. The chapter is divided into two sections, where we introduce the GDPR terms, ideas, and requirements along with the competency questions utilized here. The second part presents the necessary descriptions of the data provenance and semantic web technologies notions.

Chapter 3 — Literature Review. This chapter starts by detailing all necessary information to comprehend the systematic method employed for obtaining the current works addressing some point of the GDPR and proposing or utilizing models. Subsequently, the chapter presents a brief description of each approach collected, and it ends by showing the gaps we identified, which brings opportunities towards our contribution.

Chapter 4 — An extension to the GDPRov. This chapter presents the model related to the research objective presented in Section 1.2.3. It sheds light on the task of building such a model by minutely describing all steps involved in reaching the final model and

arguing the necessity for each decision. The final model takes a form of a small OWL2 ontology, introduced following the category of the questions as presented in the GDPR Guidance for SMEs (36).

Chapter 5 — Evaluation. This chapter presents the SPARQL queries corresponding to the competency questions over the ontologies presented in Chapter 4. This chapter demonstrates the suitability of the proposed model in answering those questions by considering running use-case examples in the literature.

Chapter 6 — Conclusion. This last chapter finishes the dissertation by summarizing the major results, consequences, and limits of the presented study. It highlights whether the results address the research question and objective. It also discusses subsequent directions regarding potential implementations and extensions through future work.

Chapter 2

Background: The GDPR Concepts and Requirements, and the Data Provenance Models

Since the contribution of this dissertation is shaped by two significant aspects (i.e., the legal and the computational aspects), this chapter is divided into two sections. Section 2.1 explains the concepts related to the legal aspects of the General Data Protection Regulation (GDPR). This first section presents a brief introduction to the GDPR, its core terminologies, and the external resources to assist with checking compliance. Hereafter, section 2.2.2.1 explores the computational aspects of this dissertation by chronicling the construction of the ideas behind ontology and provenance and by presenting the terminology and the state-of-the-art technologies utilized in the contribution and evaluation. While the first section provides a set of terms regarding WHAT we must represent, the second one presents the terms regarding HOW we must represent it.

2.1 Legal Aspect

To our knowledge, one of the deepest roots of the GDPR dates back to the 1950s, in the wake of the Second World War, when the 8º Article of the European Convention on Human Rights (ECHR) (38) established the first right to respect for a person's "privacy, his family life, home, and correspondence". Since there, alongside the expansion and popularization of information technologies, privacy concerns have arisen regarding the data a person leaves behind in these new information environments.

In the 1970s, the German Bundesrat[1] alongside the Bundestag[2], enacted the first federal data privacy act in the world, the German Bundesdatenschutzgesetz (6), which was conceived containing the principles of consent and transparency. In the next decade, the German Constitutional Court established the right to self-determination of information in the Volkszählungsurteil, the census verdict (7). This right aimed to empower German citizens to decide when and to what extent personal information could be published.

Only in the second half of the 1990s, the Council of Ministers and the European Parliament adopted the Data Protection Directive(97) (Directive 95/46/EC) by appointing a minimum set of data security norms that enables member states to enact their privacy law, following the same principles introduced by the German Courts. This legislation aimed at protecting individuals regarding the processing of their personal data.

2.1.1 General Data Protection Regulation

The General Data Protection Regulation (GDPR) (1) is a regulation in the European Union (EU) legislating over data protection and privacy for all individuals within the EU and the European Economic Area (EEA). The GDPR was sanctioned in April 2016 and became enforceable on May 2018. It was proposed to replace the Directive 95/46/EC (97), the prior EU privacy law[3].

Due to the difficulties for corporations to operate in different member states in compliance with Directive 95/46/EC, the European legislators created the GDPR. The GDPR seeks to harmonize data protection laws across the entire EU territory by uniforming all data protection and privacy data laws. The GDPR has also given individuals more control over their data and extended it to any organization (regardless of size or location) that operates the personal data of EU and EEA citizens.

To achieve this uniformization, the GDPR's lawmakers wrote 99 articles organized into 11 chapters. In this section, we pass through these chapters exploring the terms necessary to understand the contribution of this document. Within this road map, we met the basic terminologies utilized for the entire body of the GDPR, the proposed rights and obligations, and the principles that gave rise to them.

[1] The German Federal Council

[2] The German Parliament

[3] Both regulations and directives impose to all member states of the EU and the EEA with legal force rules that must be achieved. Directives, however, delegate to the member state the efforts related to the implementation. This differentiation often drives EU companies to face complex scenarios for integration, particularly regarding data protection and privacy rules.

2.1.1.1 Basic Terminology

The first chapter of the GDPR sets out the scope and the common terminology utilized in the remaining chapters. The GDPR reaches any company (also called as *controller* or *processor*) that *processes* the *personal data* of EU residents (*data subjects*), regardless of whether the processing takes place in the EU or not. From the 23 terms introduced in this chapter, Table 2.1 gathers only those we consider important for our goals. Note that the table also includes some additional terms that lie on that condition of importance.

2.1.1.2 Privacy Principles

Once we hold the core concepts of the GDPR, Chapter III goes further and introduces the principles for proper personal data processing. These seven principles sustain the rights and obligations further introduced. In this subsection, we introduce the seven principles into three groups just for ease of presentation. Table 2.2 presents them all individually along with a brief definition. Lastly, the GDPR also obligates the processor to comply with the controller's instructions regarding personal data processing according to these principles.

The first principle rules that personal data processing only happens *lawfully, fairly, and transparently* (i.e., the principles of "Lawfullness, fairness, and transparency" (1, Art. 5 (1))). In summary, a processing operation happens lawfully only under:[4] the data subject consent (1, Art. 6(1)(a)); or when the processing is necessary for: performing a contract (1, Art. 6(1)(b)), obtaining compliance with a legal obligation (1, Art. 6(1)(c)), protecting the vital interests of any natural person (1, Art. 6(1)(d)), performing a task carried out in the public interest or in the official authority exercises (1, Art. 6(1)(e)), or the legitimate controller's interests exclusively[5] (1, Art. 6(1)(f)).

The *transparency* of a data processing operation can be measured as the natural data subject can easily access and easily understand any information relating to the processing of their data (1, Rec. 39). This principle thus implies using clear and plain language to inform the individual about their data.

The second, third, and fifth principles draw the controller's limits. The first act of the controller toward personal data is the collection of the latter's data. Within the GDPR, every processing operation involving personal data is limited by a purpose, which is specific,

[4] Such as in literature, in this study, these conditional statements that make a processing operation became lawful (e.g., the data subject consent, the necessity to perform a contract, etc.) are also called by the terms *legal base* or *legal grounds*.

[5] Saving when these interests do override the interests, rights, and freedoms of data subjects.

Term	Definition
Personal data	"is any information relating to an identified, or identifiable natural person (i.e., data subject)" (1, Art. 4(1)).
Data subject	"is a natural person who can be identified, directly or indirectly, in particular by reference to an identifier" (1, Art. 4(1)).
Identifier	"can be a name, an identification number, location data, an online identifier, or one or more factors specific to the physical, physiological, genetic, mental, economic, cultural, or social identity of that natural person". (1, Art. 4(1));
Processing	"is any operation or set of operations that are performed on personal data or sets of personal data, whether or not by automated means. Some examples of processing are data collecting, recording, organization, structuring, storage, adaptation, or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction" (1, Art. 4(2)).
Controller	"is the natural or legal person, public authority, agency, or other body that, alone or jointly with others, determines the purposes and means of the processing of personal data" (1, Art. 4(7)).
Processor	"is a natural or legal person, public authority, agency, or other bodies that process personal data on behalf of the controller" (1, Art. 4(8)).
Third-Party	is "a natural or legal person, public authority, agency or body other than the data subject, controller, processor and persons who, under the direct authority of the controller or processor, are authorized to process personal data" (1, Art. 4(10)).
Data breach	is "a breach of security leading to the accidental or unlawful destruction, loss, alteration, unauthorised disclosure of, or access to, personal data transmitted, stored or otherwise processed" (1, Art. 4(12)).
Data protection officer (DPO)	It is an independent person responsible for ensuring that a controller or processor complies with the GDPR. DPOs are also responsible for advising the controller and processor on their data protection obligations, monitoring compliance with the GDPR, and acting as a point of contact for the supervisory authorities and individuals whose personal data is being processed (1, Art. 37-41).
Supervisory authority	A national "independent public authority" established by its member state and responsible for "monitoring the application of [the GDPR] and protecting the fundamental rights and freedoms of natural persons concerning the processing." These responsibilities take place by adopting the supervisory authority with investigative, corrective, authorization, and advisory powers. the power to investigate and impose fines or other sanctions for breaches of the GDPR, the power to order the suspension or limitation of processing activities, and the power to require from the controllers and processors to provide information or to cooperate in investigations (1, Arts. 51-60).

Table 2.1: The definition of the main terminology in the GDPR.

explicit, and legitimate on the step of collection ("Purpose limitation" (1, Art. 5(b))). These principles also limit the controller by the amount and the quality of the data it can collect ("Data minimization" (1, Art. 5(c))) and the period these data can be kept by it, surrounded by the purposes for which they were initially collected ("Storage limitation" (1, Art. 5(e))).

Once the GDPR has established the rules for lawful processing and drawn the limits for collecting actions in terms of personal data, the remaining principles burden the controller with the responsibility and liability to demonstrate compliance with all the seven principles(1, Art. 5(2)), including the responsibility to ensure the data is up to date and, where is fit, to take reasonable steps to erase or rectify the data (1, Art. 5(d)); and, finally, the responsibility to bear the appropriate security measures to protect personal data against unauthorized or unlawful processing and accidental loss, destruction, or damage. (1, Art. 5(f)).

2.1.1.3 The Categories of Personal Data

Along with the principles, the second chapter of the GDPR also introduces the concept of special categories of personal data. The intention to treat these terms separately resides in the fact that they require special legal bases for processing such data.

Article 9 (1) regulates the processing of special categories of personal data. This chapter thus introduces the following categories of personal data: racial or ethnic origin; political opinions; religious or philosophical beliefs; trade union membership; genetic data, biometric data [7]; data concerning health; and data concerning a natural person's sex life or sexual orientation.

2.1.1.4 Data Subject Rights

Chapter III is entirely dedicated to eliciting the main rights the data subject can exercise in each situation where a controller processes their data. Despite other chapters also assuring the data subjects further rights to seek the controller or other agents to ensure whole control of their data, the rights introduced in this chapter do represent the main ones in the GDPR regarding their data processing.

There are roughly, six rights the data subject can exercise regarding their data. They are introduced and summarized by Table 2.3. Finally, the GDPR imposes some constraints on the exercising of these rights. The data subject is free of charge on requesting any of these

[7] When the processing of such data is to identify a natural person uniquely.

Principle	Definition
Lawfulness, fairness, and transparency	Establishes that personal data processing only happens lawfully, fairly, and transparently. A processing operation is transparent when the natural data subject can easily access and easily understand (which implies the use of clear and plain language) any information related to the processing of their data;
Purpose limitation	This principle requires, by the controller, the declaration of the purposes for collecting personal data. This data collection purpose must be in such a manner specific, explicit, and legitimate so that further processing with distinguishable purposes can be identified and thus do not occur[6];
Data minimization	The amount and the quality of collected data are restricted to what is necessary for the purposes for which they are processed;
Accuracy	The controller is responsible for ensuring, where is fit, the up to date of personal data; and, once it is out of date, for taking reasonable steps to erase or rectify it without delay.
Storage limitation	This principle mandates the personal data collected is kept only for the period it is necessary, according to the purposes for which they were collected;
Integrity and confidentiality	The controller is responsible for ensuring appropriate security of the personal data, which includes protection against unauthorized or unlawful processing and accidental loss, destruction, or damage. For this, the controller must use appropriate technical or organizational measures;
Accountability	Any processing of personal data carried out by the controller or on the controller's behalf (e.g., the data processing carried out by the processor on behalf of the controller) is under the responsibility and liability of the controller that has determined the purposes and means for this processing.

Table 2.2: The seven principles introduced by the GDPR.

rights, and controllers are required to respond to these requests within one month unless these requests are particularly complex, in which case the response time may be extended to two months.

2.1.1.5 Controller and Processor Obligations

GDPR entitles Chapter IV as "Controller and processor" due to set out the obligations of controllers and processors with regard to the processing of personal data, including the obligation to implement appropriate technical and organizational measures to protect personal data, the obligation to respect the rights of data subjects, and the obligation to notify supervisory authorities of personal data breaches.

Right	Definition
Right of access by the data subject	This right empowers the data subject to get the confirmation of whether or not their data is being processed, and if so, the accessing of these data and any information available about the processing of their data, including information about its rights (1, Art. 15).
Right to rectification	Under this right, the data subject can demand from the controller the correction if their data is inaccurate, or completion, if they are incomplete (1, Art. 16).
Right to erasure	The data subject can request the controller to delete the data concerning him (this right is also referenced by "Right to be forgotten") (1, Art. 16).
Right to object or restrict processing	The data subject can reject the processing involving their data, or request that the processing of their data be restricted (1, Arts. 18, 21).
Right to data portability	The data subject request to receive from the controller their data; or that the controller transfers these data to another controller (1, Art. 20).
Right to reject automated decisions with legal effects	Under this right, the data subject can is not obligated to be subject to a decision based solely on automated processing (1, Art. 20).

Table 2.3: The data subject rights defined by the GDPR.

Presenting all these obligations in detail is challenging and goes beyond the purposes of this work. For the sake of introduction, Table 2.4 synthesizes all obligations of the controllers and processors established by the GDPR into six obligations. This list is not mandatory or exhaustive and might variate depending on the author in the literature. However, it certainly covers the main obligations established by the GDPR.

2.1.1.6 Summary

The remaining chapters lay down rules to regulate some specificities and details that support the implementation of the rights and obligations elicited up to this point, including the rules, conditions, and mechanisms for the permitted transferring of personal data to countries or territories outside the EU or the EEA (1, Arts. 44-50); the procedures for the cooperation and consistency of supervisory authorities in order to ensure consistent application of the GDPR across the EU (including the procedures for the resolution of disputes between supervisory authorities) (1, Arts. 63 to 67); the procedures for the appeals and judicial remedies available to controllers, processors, and data subjects (including the individuals' right to an effective judicial remedy against a supervisory authority's decision; and the controllers and

Obligation	Definition
Take technical and organizational measures.	This obligation sets out the general requirement for controllers to implement appropriate technical and organizational measures to ensure and demonstrate that the processing of personal data is performed in accordance with the GDPR, including data protection by design and by default, taking into account the risks to individuals when determining these appropriate technical and organizational measures, and aiming to protect personal data against unauthorized or unlawful processing, accidental loss, destruction, or damage (i.e., personal data breach) (1, Art. 24-26, 32).
Designate and cooperate with supervisory authorities.	This obligation synthesizes three articles that require controllers to designate a representative in the EU if they are not established in the EU (1, Art. 27), to enter into a contract with the processors that carried out his processing (1, Art. 28), to cooperate with supervisory authorities and to provide them with any information they require to carry out their duties (1, Art. 29).
Keep records of processing activities.	This obligation requires the controllers to maintain records of their processing activities, including the purposes of the processing, the categories of personal data involved, and the categories of recipients to whom the personal data has been or will be disclosed (1, Art. 30).
Cooperate, notify, and inform in case of a personal data breach.	Under this obligation, the controllers shall cooperate with and notify the supervisory authorities and processors in the event of a personal data breach, especially when such an event poses a risk to individuals. In this last case, the controller shall inform individuals about personal data breaches that pose a high risk to their rights and freedoms (1, Arts. 31, 33, 34).
Take data protection impact assessment and prior consultation.	Articles 35 and 36 cover the requirement for controllers to carry out a data protection impact assessment (DPIA) in certain circumstances, and to consult with the supervisory authority prior to carrying out processing that is likely to result in a high risk to the rights and freedoms of individuals (1, Arts. 35, 36).
Designate a data protection officer.	In cases when the core activities of the controller or processor consist of processing operations that require regular and systematic monitoring of data subjects on a large scale or when a public authority or body carries out the processing, the controller and processor must designate a DPO. (1, Arts. 37).

Table 2.4: The controller and processor obligations according to the GDPR.

processors' right to appeal against decisions of supervisory authorities) (1, Arts. 68-72), and so on.

Nevertheless, the terms introduced in Tables 2.1, 2.2, 2.3, and 2.4 are enough for a complete comprehension of the basics of the GDPR and they summarizes the sufficient notions for the understanding of this dissertation.

2.1.2 The GDPR Guidance For SMEs

Even though the controllers hold these basic notions of the GDPR, it is still necessary to translate these legal terminologies and obligations into daily practical requirements. Within software engineering, translation into software requirements is challenging, error-prone, and non-trivial tasks (81; 54). The adaptation of legal requirements into software requirements is often too abstract, and thus may lead to interpretations and inferences that are inconsistent (15; 81; 66; 74). As an example, although the GDPR does not clearly define "privacy by design", it requires the controllers to implement "data protection by design and by default" (1, Art. 25) (15; 66) [8]. Moreover, how to measure the adequacy level of protection (1, Art. 45) of personal data (74)? What does the obligation of taking data protection impact assessment mean for small and medium enterprises (SME)? How can it be evaluated in order to certify whether the enterprise complies with the GDPR?

In conclusion, Ayla-Riviera and Pasquale (15) suggest the necessity of guidance that clarifies these terms and requirements at a level that capable controllers self-assess their compliance. Prior publications concerning this matter have supported and contributed to this suggestion (25; 42; 85).

Indeed, supervisory authorities have traditionally provided guidance to support the controllers to comply with the GDPR [9]. For example, the Spanish Data Protection Agency (AEPD)[10] has provided training to controllers on topics such as data protection by design and default, data protection impact assessments, and the rights of data subjects.

Another example is the provision of a checklist to help organizations to ensure GDPR

[8] Ayala-Rivera and Pasquale (15) shed lights on a linguistic issue: lawyers makers and courts endeavor to describe things using legal terminologies. In that process, they propose and enrich a common legal vocabulary capable of describing such things. At the same time, software engineers and IT stakeholders have also evolve their language to describe the things they have interested. This latter vocabulary is composed by different symbols and grammars, it goes beyond the common human languages (e.g., English, Portuguese) touching formal languages and rules that establishes the relation between the symbols. This linguistic problem concerns about the translation the terms between different realms of knowledge, the impossibility of existing final methodology that ensures identical translation of terms, and the difficulties to measure, compare, and evaluate different translations.

[9] Although the GDPR does not clearly burden supervisory authorities with this task (despite some points (1, Art. 57(j), (k), (l), (p)) might support this), historically, this role has been played by regulatory bodies, and it is a common practice to ensure compliance with laws and regulations. For example, the "Guide to Data Protection" (62) was published by the Information Commissioner's Office (ICO) in the United Kingdom as guidance to enterprises on how to comply with the DPA (97). Additionally, The Article 29 Working Party (now the European Data Protection Board (1, Art. 68)), which has served as the advisory body and has provided guidance on the elaboration and application of the GDPR, states that supervisory authorities should "provide guidance and support to controllers and processors on the certification process, including how to prepare for certification, how to choose a certification body, and how to maintain certification over time" (44).

[10] The supervisory authority in Spain

Category	ID	Title	Competency Question
General	CQ8	Retention period	For each category of personal data, list the period for which the data will be retained e.g. one month? one year? As a general rule, data must be retained for no longer than is necessary for the purpose for which it was collected in the first place.
	CQ9	Action required to be GDPR compliant?	Identify actions that are required to ensure all personal data processing operations are GDPR compliant. E.g. this may include deleting data where there is no further purpose for retention.
Personal Data	CQ11	Retrospective Consent	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?
Rights	CQ17	Subject Access Requests (SARs) Response Time	Is your organization able to respond to SARs within one month?
	CQ20	Right to restriction of processing	Are there controls and procedures in place to halt the processing of personal data where an individual has, on valid grounds, sought the restriction of processing?
	CQ21	Right to object to processing	Are individuals told about their right to object to certain types of processing, such as direct marketing, or where the legal basis of the processing is legitimate interest or necessary for a task carried out in the public interest?
	CQ25	Restrictions to data subject rights	Have the circumstances been documented in which an individual's data protection rights may be lawfully restricted? Note: the Irish Data Protection Bill will set out further details on the implementation of Article 23.
Accuracy Retention	CQ28	Accuracy	Are procedures in place to ensure personal data is kept up to date and accurate and where a correction is required, the necessary changes are made without delay?
	CQ29	Retention	Are retention policies and procedures in place to ensure data is held for no longer than is necessary for the purposes for which it was collected?
	CQ30	Retention Legal Obligations	Is your business subject to other rules that require a minimum retention period (e.g. medical records/tax records)?
	CQ32	Duplication of records	Are procedures in place to ensure that there is no unnecessary or unregulated duplication of records?
Transparency	CQ33	Transparency to customers and employees	Are service users/employees fully informed of how you use their data in a concise, transparent, intelligible, and easily accessible form using clear and plain language?
	CQ37	Provide information when engaging	When engaging with individuals, such as when providing a service, sale of a good, or CCTV monitoring, are procedures in place to proactively inform individuals of their GDPR rights?
	CQ38	Provide information on facilitating rights	Is the information on how the organization facilitates individuals exercising their GDPR rights published in an easily accessible and readable format?

Continued on the next page

compliance. The Irish Data Protection Commission (DPC)[11] has published a plethora of documents approaching specific topics of the GDPR. One of these intends to assist Small to Medium Enterprises (SMEs) in mapping the personal data they have collected, the legal basis

[11]The supervisory authority in Ireland

Category	ID	Title	Competency Question
Controller Obligations	CQ39	Supplier Agreements	Have agreements with suppliers and other third parties processing personal data on your behalf been reviewed to ensure all appropriate data protection requirements are included?
	CQ40	Data Protection Officers	Do you need to appoint a DPO as per Article 37 of the GDPR?
	CQ41	Reasons for not having a DPO	If it is decided that a DPO is not required, have you documented the reasons why?
	CQ42	Escalation procedures	Where a DPO is appointed, are escalation and reporting lines in place? Are these procedures documented?
Data Security	CQ47	Documented Security Program	Is there a documented security program that specifies the technical, administrative, and physical safeguards for personal data?
	CQ48	Resolving security related issues	Is there a documented process for resolving security-related complaints and issues?
	CQ49	Designated individual for security	Is there a designated individual responsible for preventing and investigating security breaches?
	CQ50	Encryption	Are industry-standard encryption technologies employed for transferring, storing, and receiving individuals' sensitive personal information?
	CQ51	Removing information	Is personal information systematically destroyed, erased, or anonymized when it is no longer legally required to be retained?
	CQ52	Restoring access	Can access to personal data be restored in a timely manner in the event of a physical or technical incident?
Data Breach	CQ54	Regular reviews	Are plans and procedures regularly reviewed?
	CQ57	Documentation of data breaches	Are all data breaches fully documented?
	CQ58	Co-operation procedures for data breach	Are there cooperation procedures in place between data controllers, suppliers, and other partners to deal with data breaches?
International Data Transfer	CQ63	Transfer Details	Are all transfers listed - including answers to the previous questions (e.g. the nature of the data, the purpose of the processing, from which country the data is exported and which country receives the data, and who the recipient of the transfer is?)
	CQ64	Legality of international transfers	Is there a legal basis for the transfer, e.g. EU Commission adequacy decision; or standard contractual clauses? Are these bases documented?
	CQ65	Transparency	Are data subjects fully informed about any intended international transfers of their personal data?

Table 2.5: Competency Questions (Continued)

for them, the processing carried out upon these data, the retention period for each category of data, and so forth. DPC titled this document as Guidance for SMEs (36). Among different sections in the Guidance for SMEs, we are more interested in the last one, "GDPR Readiness Checklist Tools", which provides a checklist with questions organized into different areas concerning GDPR terms.

In this research, we utilize part of the questions introduced in the Guidance for SMEs (36) as competency questions (CQ) to fulfill the needs elicited in the Research Methodology

section, in Chapter 1.Indeed, a prior work (85) has utilized these questions following a similar methodology we apply in this study, addressing 32 of the 65 obligations listed in the 2019 version of the Guidance and leaving 33 uncovered. Table 2.5 lists these uncovered questions. Of those 33, we address four (CQ8, CQ28, CQ29, and CQ51) in Chapter 4, while the remaining 29 are organized into a five-category map of provenance-amenable topics that suggests systematic pathways for future extensions.

2.2 Computational aspect

The first section of this chapter provides a set of terms regarding *WHAT* we must represent. Now, we need to understand the terms related to *HOW* we must represent. These latter terms belong to the computational aspect of the GDPR. What are thus the concepts and technologies the computer scientists have developed that can assist us in achieving the RO?

This research aims to find a *model* to *represent, query, and validate* the information used in answering the questions presented in the Guidance for SMEs (36). By seeking to represent a body of information, Computer Science research communities have borrowed the concept of *ontology* and provided a set of tools that make it possible to *represent, query, and validate* a terminology in a machine-readable format.

For the purpose of this work, *model* and *ontology* are interchangeable concepts. Atkinson et al. (13) demonstrate the reasoning for this position. By analyzing how the literature has commonly and informally distinguished "model" from "ontology" (i.e., considering their purposes or characteristics) and taking common definitions for both terms, they have concluded that we can use both terms interchangeably for those kinds of models in which the "main purpose is to explicitly and clearly describe the concepts in a particular domain of interest." This model is often referred to in the literature as the "domain model." That definition fits well into the purposes of this study[12].

Considering thus the definition of *ontology*, in the computer science contexts, it is a formal and explicit specification of a shared, "abstracted and simplified view of the world we wish to represent for some purpose" (59; 79). This representation includes "the objects, concepts, and entities presumed to exist in some area of interest and the relationships that hold them" (59; 51).

As an example, suppose an ontology that represents part of what a university is. By cap-

[12]The non-distinction of both concepts is worth discussing due to the results we have collected from the literature in Chapter 3, in which both terms have been used interchangeably.

turing the concepts and relationships within the context of the university, we can describe an ontology to represent the university's structure and organization. This can include entities such as departments, faculties, students, staff, courses, research projects, etc. This ontology can also represent the relationships between these entities, such as student enrollments, course scheduling, research management, etc. Other ontologies can yet exist in parallel, aiming to represent the concepts and relations outside the university premises, such as their legal or financial commitment to the government or research sponsors. In all cases, the ontology is an abstraction according to some purpose. Once this ontology is represented in a machine-readable format, it is possible to draw conclusions, query, and validate assertions about the entities, relations, and individual information modeled by the ontology.

Since this work contributes by constructing an ontology, the next step consists of introducing the technologies that make it possible to create and utilize an ontology according to a machine-readable format so that we meet our goals, which is precisely what Section 2.2.1 does.

2.2.1 The Semantic Web Technologies

Such as the ancient voices in the oldest books, Sir Tim Berners-Lee, in his famous publication in a 2001 Scientific American paper (18) namely "the Semantic Web", presaged a nowadays-common use case of the internet:

> "By only interacting with their own device, people would be able to consume, record, and indirectly touch a plethora of resources, agents, and information available on the internet – things they would have never imagined." (18)

This entire process is thanks to the *Semantic Web*, which extends and structures the web of content so that software agents can roam among the web documents to perform advanced tasks. To achieve this goal, ontology is a crucial component, as it provides the structure and the vocabulary for this information on the web. Semantic Web contributors thus have made efforts to design standardized and machine-readable formats for representing these ontologies. These efforts have resulted in technologies that are highly interoperable and reusable, at the same time, allow querying and reasoning over the data that the ontological parts know.

The World Wide Web Consortium (W3C)[13] is the pivot contributor to the Semantic Web. It has developed standards and technologies to make it possible to implement the semantic

[13] https://www.w3.org/Consortium/

web. One of the basic standards is the *Resource Description Framework* (RDF), which is a data model for describing resources on the internet. RDF represents a piece of information as a triple: subject-predicate-object. In such a triple, the predicate expresses the existing relationship (called property by the RDF vocabulary) between the subject (an RDF resource) and the object (resource or literal). For example, the information that "Socrates is mortal" can be represented as an RDF triple as ⟨Socrates, *is*, mortal⟩.

2.2.1.1 RDF Serialization Syntax

To make an RDF triple machine-readable, the W3C has developed different syntax that serializes the set of statements according to different purposes. The RDF/XML[14] is the original syntax for RDF and is based on the XML markup language[15]. It suits well with existing XML tools and technologies. The Terse RDF Triple Language[16] (Turtle) is designed to be easy to read and write by humans. The JSON-LD[17] encodes the RDF triple in the JSON data format, which is generally compatible with most existing web ecosystems, and it is easy to consume by web developers. The previous example can be represented using Turtle syntax as shown in Listing 1.

```
1 <http://example.com/#Socrates> <http://example.com/#is>
  ↪ <http://example.com/#mortal> .
```

Listing 1: Example of a triple in Turtle format.

Within RDF, each part of the triple receives a unique global identifier, named Uniform Resource Identifiers (URIs) [18], which allows for sharing, reusing, and interoperability between different ontologies, since each part of the triple can be individually identified.

The Turtle format also supports a more compact representation by allowing the notation for a URI namespace as the shorthand prefixes. These prefixes must be defined at the beginning of the document. The shorthand prefix is defined in a line initiated by the reserved word **@prefix**, followed by the alphanumerical chain terminated by the colon, that carries the prefix identifier, and ended by a string of characters enclosed in angle brackets and a dot. The URI in Listing 1 can be shorthand defined like `@prefix ex: <http://example.com/> .`.

[14] https://www.w3.org/TR/rdf-syntax-grammar/
[15] https://www.w3.org/TR/2008/REC-xml-20081126/
[16] https://www.w3.org/TR/turtle/
[17] https://www.w3.org/TR/json-ld/
[18] http://www.faqs.org/rfcs/rfc2396.html

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix ex: <http://example.com#> .

ex:Mortal rdf:type owl:Class ;
          rdfs:comment "The class of those things that are subject to
 ↪  death"@en ;
          rdfs:label "Mortal" .

ex:Human rdf:type owl:Class ;
         rdfs:subClassOf ex:Mortal ;
         rdfs:comment "Taxonomically Homo sapiens, a term that derives
 ↪  from the Latin 'wise man', also known as person, people or man"@en ;
         rdfs:label "Human" .

ex:Socrates rdf:type owl:NamedIndividual, ex:Human .
            rdfs:comment "Greek philosopher from Athens who is credited
 ↪  as the founder of Western philosophy"@en ;
            rdfs:seeAlso <https://en.wikipedia.org/wiki/Socrates> .

# Derive that Socrates is Mortal
# ex:Socrates rdf:type :Mortal .
```

Listing 2: Expressing the classical syllogism “All humans are mortal; Socrates is human; Then, Socrates is mortal” in RDF triples using OWL and RDFS in turtle format.

To use an abbreviated prefix, one needs to inform the identifier and name of the desired ontology concept, such as **ex:Socrates**.

2.2.1.2 RDF Vocabulary to Describe Ontologies

Upon the basic RDF data model, further vocabularies were built to provide additional functionality in describing RDF triples. Among the most important, we highlight the RDF Schema (RDFS)[19] and the Web Ontology Language (OWL)[20].

Among W3C standards we highlight *RDF Schema (RDFS)*[21] and the *Web Ontology Language (OWL)*[22], both used to define structured, machine-readable information and the relationships between data. OWL is an expressive language for creating ontologies: it lets us

[19] https://www.w3.org/TR/rdf-schema/
[20] https://www.w3.org/TR/owl-features/
[21] https://www.w3.org/TR/rdf-schema/
[22] https://www.w3.org/TR/owl-features/

```
PREFIX ex: <https://example.com/>

SELECT ?person
WHERE {
  ?person rdf:type ?class .
  ?class rdfs:subClassOf* ex:Mortal .
}
```

(A) The SPARQL query.

```
<https://example.com/#Socrates>
```

(B) The SPARQL result.

Listing 3: An example of a SPARQL query (A) and the query result (B) over a RDF dataset where the classical argument "All humans are mortal; Socrates is human; Then, Socrates is mortal" are instantiated. This query are seeking for any instance of subclass of **ex:Mortal**.

define classes, properties, and individuals and provides rich operators for building complex class descriptions and for stating relationships between classes and individuals.

OWL[23] goes beyond RDFS concepts and provides a rich vocabulary capable of detailing the relationship between classes and properties for an ontology. It allows the specification of restrictions concerning the values of a property, the equivalence or difference existing between two concepts, etc.

Listing 2 shows these ideas in practice. Lines 6–13 declare the classes **ex:Mortal** and **ex:Human** and encode the first premise -— "All humans are mortal"—by asserting that **ex:Human rdfs:subClassOf ex:Mortal**. Line 15 adds the second premise, "Socrates is human", via **ex:Socrates rdf:type ex:Human**. We do not need to write the conclusion triple "Socrates is mortal", because OWL and RDFS let a reasoner infer the conclusion "Socrates is mortal" from the first two axioms together. In short, RDFS and OWL let us declare concepts through **rdf:type** as either **owl:Class** or **owl:NamedIndividual** and enrich them with a **rdfs:label**, a detailed **rdfs:comment**, or additional links via **rdfs:seeAlso**.

2.2.1.3 SPARQL: An Ontology Query Language

Once we have a tool to represent an ontology, we need an RDF query language. For this purpose, the W3C developed *SPARQL*[24], a language for retrieving specific triples from RDF datasets. Its syntax is similar to SQL, yet it was designed for graph-like datasets built with

[23] http://www.w3.org/2002/07/owl#
[24] https://www.w3.org/TR/sparql11-query/

triples, and each query has two parts: first we declare the variables we want to retrieve; then we connect those variables according to our interests.

A variable is referred to by the symbol "`?`" followed by an identifier, e.g. `?person`. We write the variables we want to retrieve on the first line after the reserved keyword `SELECT`. The second part of the query lives in a `WHERE` block, where we connect the variables to create a pattern of data; every occurrence that matches this pattern is retrieved. The W3C proposes a myriad of keywords and options to shape such patterns. Listing 3(A) encodes the question "how is mortal?" into a SPARQL query. This query retrieves all instances of **`ex:Mortal`** (including instances of subclasses); when run over the triples introduced in Listing 2, it produces the result depicted in Listing 3(B).

2.2.1.4 Ontology Validation Language

Finally, to achieve the goals regarding the validation of the data, the Shapes Constraint Language[25] (SHACL) is recommended. SHACL is a W3C language that provides the capability to specify constraints (shape graph) to validate information described in a set of RDF triples (data graph). The shape graph is described in terms of RDF triples; thus, it can also be serialized using RDF languages (e.g, Turtle, JSON-LD).

After the validation process, the SHACL reports the result of whether the data graph conforms or not with the specified shape graph. In the negative case, attribute violations are included in the answer.

To make SHACL easier to understand, take as the data graph the example presented in Listing 2, extended by the triples in Listing 4(C). Suppose we need to specify that **`onto:Person`** instances must have exactly one value for the property **`onto:name`**. Suppose also that this value must be a string, and the value of the property **`onto:author`** must be an instance of **`onto:Book`**. These constraints can be implemented through SHACL as specified in Listing 4(A).

Listing 4(B) shows the validation result graph after the triples that invalidate the constraints elicited above were added to the data graph. For this example, we add the following triples:

[25] `https://www.w3.org/TR/shacl/`

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix sh: <http://www.w3.org/ns/shacl#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix onto: <http://example.org/ontology/#> .
@prefix ex: <http://example.org/instances/#> .
```

```
ex:PersonShape
  a sh:NodeShape ;
  sh:targetClass onto:Person ;
  rdfs:label "Person" ;
  sh:property ex:NameProperty ,
          ex:authorProperty .

ex:NameProperty
  sh:path onto:name ;
  sh:name "Name";
  sh:datatype xsd:string ;
  sh:maxCount 1 ;
  sh:minCount 1 .

ex:authorProperty
  sh:path onto:author ;
  sh:name "Authorship" ;
  sh:class onto:Book .
```

(A) Shade graph definition

```
_:report a sh:ValidationReport ;
    sh:conforms false ;
    sh:result [
        rdf:type sh:ValidationResult ;
        sh:resultSeverity sh:Violation ;
        sh:sourceConstraintComponent
↪   sh:MaxCountConstraintComponent ;
        sh:sourceShape ex:NameProperty ;
        sh:focusNode ex:JRRTolkien ;
        sh:resultPath onto:name ;
        sh:resultMessage "More than 1
↪   values" ;
    ], [
        rdf:type sh:ValidationResult ;
        sh:resultSeverity sh:Violation ;
        sh:sourceConstraintComponent
↪   sh:ClassConstraintComponent ;
        sh:sourceShape ex:AuthorProperty ;
        sh:focusNode ex:JaneAusten ;
        sh:resultPath onto:author ;
        sh:value ex:JRRTolkien ;
    ] .
```

(B) Validation result graph

```
ex:JRRTolkien onto:name "John Ronald Reuel Tolkien" .
ex:JaneAusten rdf:type onto:Person .
ex:JaneAusten onto:author ex:JRRTolkien .
```

(C) The remaining triples of the data graph, considering those in Listing 2

Listing 4: A simple SHACL example.

2.2.2 Data Provenance

Since the GDPR rules the processing of personal data, it is essential to represent the chain of transformation each piece of data was submitted to along the period of its life, occurring from the collection instant at the origin until the time the data is no more necessary for the purpose it was initially collected and thus deleted. Indeed, Article 30, as Table 2.4 illustrates, obligates controllers to record the logs concerning the processing activities, including information on when and where each procedure ran and how and who was involved.

"The origins of something, its history, derivation, and passage through its various owners" is the definition of *provenance* according to the Oxford English Dictionary. Some research communities have borrowed this term and developed it to express the *provenance of data*. Data provenance was born in scientific workflow management to help scientists to assure the reproducibility and reuse of scientific experiments. Due to ensuring reliability, and trustworthiness, data provenance suits well in meeting data privacy laws provisions related to the principle of transparency and accountability, in particular, the stated in Article 30.

In this context, when personal data is collected, processed, and disclosed, data provenance information can be used to track where the data came from, who has accessed it, what it has been used for, and so on. Through data provenance, the data subjects can be aware of what data is necessary to process and keep stored and what data can be deleted or otherwise not used.

Alongside Article 30, Article 35 mandates the specification of some envisaged processing operations so that the controller can obtain an estimate of the impact of these operations on the protection of personal data. Indeed, Pandit (85) has identified this necessity and addressed it by promoting the use of prospective provenance.

In the data provenance literature (49) , some authors have discriminated the concept of provenance into two: *prospective provenance* is the set of information that specifies the workflow that must be done to produce some data. It is generally used as the analogy of a recipe, in which, prior to the final product, we already have the entire information involved in producing it. On the other side, *retrospective provenance* is looking-backward information, chronicling the activities and the runtime environment states during the production of some data, including the resources and agents involved. It is generally instantiated by a set of execution logs, which carries the snapshots of the execution along the time.

2.2.2.1 PROV: Representing Retrospective Provenance

To describe the data provenance information, W3C has developed the PROV[26], which is a family of documents that provides standards, vocabularies, and serializations of the provenance of resources in the web context.

Within PROV, provenance is seen as a "record that describes the people, institutions, entities, and activities involved in producing, influencing, or delivering a piece of data." To represent that, the PROV data model (PROV-DM) counts with many concepts divided into six

[26]http://www.w3.org/TR/prov-overview/

Concept	Definition
`prov:Entity`	A class to represent an object with some fixed attributes, whether it's physical, digital, conceptual, real, imaginary, or other.
`prov:Activity`	A class that represents 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.
`prov:Agent`	A class to represent something that bears some form of responsibility for an activity taking place, for the existence of an entity, or for another agent's activity.
`prov:wasGeneratedBy`	A relation for representing the completion of production of a new entity by an activity. The new entity begins to exist and thus becomes available just after this generation.
`prov:Used`	A relation for representing that an activity uses an entity.
`prov:wasInformedBy`	A relation representing the exchange of some unspecified entity by two activities, one activity using some entity generated by the other.
`prov:wasDervivedFrom`	A relation between two entities, representing a construction, an update, or a transformation of an entity into another.
`prov:wasAttributedTo`	A relation used to represent the case where an agent is the author or responsible party for an entity.
`prov:wasAssociatedWith`	A relation representing the responsibility of an agent for an activity, indicating that the agent had a role in the activity.
`prov:actedOBehalfOf`	A relation denoting the assignment of authority and responsibility to an agent to carry out a specific activity as a delegate or representative, while the agent it acts on behalf of retains some responsibility for the outcome of the delegated work.

Table 2.6: PROV-DM core classes and relations

parts. The core provenance concepts, according to PROV, are composed of three classes and seven properties (relationships). Table 2.6 contains the definitions of these core concepts.

Taken together, these classes and relations induce a *provenance graph*: a directed graph whose nodes are the entities, activities, and agents involved in handling a piece of data, and whose edges are the provenance relations that connect them (for instance, an activity that `prov:used` an entity, or an entity that `prov:wasGeneratedBy` an activity). Reading such a graph backward from a given data item reconstructs its *lineage* (the chain of activities and agents that produced it), whereas reading it forward traces how that item is subsequently used. This graph-based view underlies the remainder of the dissertation: representing a

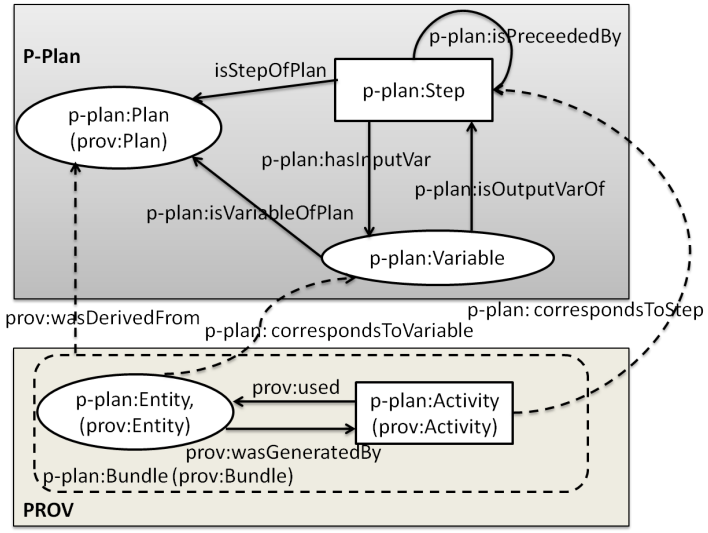

Figure 2.1: P-Plan components and their relationship with PROV (50)

competency question amounts to describing a characteristic sub-structure of the provenance graph, and answering it amounts to querying the graph for the presence (or absence) of that sub-structure.

2.2.3 P-Plan: Representing Prospective Provenance

As already introduced, the prospective provenance are concerned to prescribe what should be occur with some data. It can be described through the P-PLAN Ontology[27], which extends (3) framework but specifies to model plans that guide the execution of scientific processes, enhancing transparency and accountability in scientific research.

P-PLAN consists of several key components designed to represent different aspects of a scientific plan. The core elements include **`p-plan:Plan`**, **`p-plan:Step`**, and **`p-plan:Variable`**, where a **`p-plan:Plan`** encompasses multiple **`p-plan:Steps`**, and each **`p-plan:Step`** can have associated **`p-plan:Variables`** that specify inputs, outputs, and parameters. These elements are interconnected with properties such as **`p-plan:hasStep`**, **`p-plan:isStepOfPlan`**, and **`p-plan:hasVariable`**, facilitating a detailed and hierarchical representation of the workflow. Figure 2.1 depicts the core elements of P-PLAN.

2.2.4 SSN: Representing Sensors and Observations

The vocabularies presented so far describe the provenance of data and of plans, but none of them provides terms for representing the entities that observe the state of the world and

[27] Available at `http://purl.org/net/p-plan`

connect events to the conditions that trigger them. The *Semantic Sensor Network* (SSN) ontology[28], a W3C Recommendation, fills this gap by providing a framework for representing sensors and their observations, along with the procedures involved, the features of interest, and the properties being observed. Its companion and lightweight core, the *Sensor, Observation, Sample, and Actuator* (SOSA) ontology, captures the essential observation concepts in a self-contained way: `sosa:Sensor` represents the entity that observes a property, `sosa:ObservableProperty` represents the property being observed, and the `sosa:observes` relation links a sensor to the properties it can sense.

By framing an observation as the causal connection between the state of a property and the occurrence of an event, SSN supplies the vocabulary with which this dissertation represents the causal relationship between the fulfilment of a data-collection purpose and the triggering of a deletion or anonymization procedure, the concern of CQ29 (Chapter 4).

2.2.5 GDPRov: A Provenance Model for the GDPR

The vocabularies presented so far are domain-independent: PROV describes entities, activities, and agents in the abstract, P-Plan describes plans and steps without any commitment to a particular application, and SSN describes sensors and their observations. Building a compliance-checking model on top of them requires specializing these generic concepts into the concrete vocabulary of data protection: controllers, data subjects, personal data, consent, and the procedures of collection, rectification, storage, and deletion. GDPRov (92; 85) is precisely such a specialization. Proposed by Pandit, it is a data provenance ontology that represents the provenance of consent and of the activities performed over personal data, with the explicit aim of documenting and demonstrating GDPR compliance. It is also the ontology this dissertation extends, which is why we describe its organizing principles here in some detail.

GDPRov is assembled by reusing three existing vocabularies. It extends P-Plan (and, through it, PROV-O) to describe the provenance of activities over personal data, and it incorporates the Open Digital Rights Language (ODRL)[29] to describe consent and the permissions and obligations attached to processing. From P-Plan and PROV it inherits the central distinction introduced above: every relevant procedure can be described *before* execution, as a plan (prospective, ex-ante), and *after* execution, as a record of what actually happened (retrospective, ex-post).

[28] Publicly available at `https://www.w3.org/TR/vocab-ssn/`

[29] ODRL is a W3C Recommendation for expressing policies (permissions, prohibitions, and obligations) over digital assets. See `https://www.w3.org/TR/odrl-model/`.

Concept	Definition
`gdprov:DataStep`	The plan for a procedure over personal data, describing it *before* execution (prospective). It specializes `p-plan:Step`; subclasses include `gdprov:DataCollectionStep` and `gdprov:DataDeletionStep`.
`gdprov:DataActivity`	The record of a procedure over personal data that has actually run (retrospective). It specializes `prov:Activity`; subclasses include `gdprov:DataCollectionActivity` and `gdprov:DataDeletionActivity`.
`gdprov:Data`, `gdprov:PersonalData`	Data as foreseen in a plan (prospective). `gdprov:PersonalData` is the subclass whose instances can identify a data subject; categories of personal data are declared as its subclasses.
`gdprov:DataEntity`, `gdprov:PersonalDataEntity`	A concrete data item that took part in an executed activity (retrospective), the ex-post counterpart of `gdprov:Data` and `gdprov:PersonalData`.
`gdprov:Controller`, `gdprov:DataSubject`, `gdprov:ThirdParty`	The agents acting in the data lifecycle, specializing `prov:Agent` to the roles the GDPR defines.

Table 2.7: GDPRov core organizing concepts. The prefixes `gdprov:`, `p-plan:`, and `prov:` refer to the GDPRov, P-Plan, and PROV vocabularies, respectively.

This dual view is the backbone of GDPRov, and it is realized systematically through parallel families of classes. Prospective plans are modeled with *Step* classes, which specialize `p-plan:Step`; retrospective executions are modeled with *Activity* classes, which specialize `prov:Activity`. The plan to collect data, for example, is a `gdprov:DataCollectionStep`, whereas an actual collection event is a `gdprov:DataCollectionActivity`. The data itself is represented in two forms as well: `gdprov:Data` (and its subclass `gdprov:PersonalData`) denotes data as foreseen in a plan, while `gdprov:DataEntity` (and `gdprov:PersonalDataEntity`) denotes a concrete data item that took part in an executed activity. This parallelism lets a controller state, ahead of time, which procedures are planned over which categories of personal data, and later record which procedures were in fact carried out (both being necessary to demonstrate compliance).

With these building blocks, GDPRov covers the lifecycle of personal data (its collection, usage, storage, sharing, rectification, and deletion or anonymization), each expressed in both its prospective and its retrospective form. Agents are specialized as well, into `gdprov:Controller`, `gdprov:DataSubject`, and `gdprov:ThirdParty`, mirroring the roles the GDPR defines (Section 2.1.1). Table 2.7 summarizes the core organizing concepts of GDPRov that recur throughout this dissertation.

In its published form, GDPRov was evaluated against the compliance questions of the GDPR Guidance for SMEs and shown to answer part of them, leaving others open, the gap this dissertation addresses (Chapter 4). The Step/Activity and Data/DataEntity dualities are therefore a prerequisite for the extensions we introduce: each newly covered competency question is addressed by adding, in this same systematic fashion, the prospective and retrospective classes it requires.

2.3 Chapter Summary

This chapter introduced the two pillars of this dissertation: the legal and the computational aspects of GDPR compliance. The first section presented the terminology the GDPR establishes to frame data protection – personal data, data subjects, controllers, and processors – together with the seven principles of Article 5, the rights granted to data subjects, and the obligations imposed on controllers and processors. It also introduced the GDPR Guidance for SMEs (36), the official checklist of compliance questions whose coverage this work extends. The second section presented the computational side: the Semantic Web stack (RDF, OWL, SPARQL) and the provenance vocabularies that ground our model. Among these, the PROV data model provides the retrospective representation of activities, P-Plan the prospective representation of plans, SSN the representation of sensors and observations, and GDPRov the specialization of these generic vocabularies into the concepts of data protection. The central distinction carried throughout the dissertation is GDPRov's Step/Activity duality, which lets every procedure be described both before execution (as a plan) and after execution (as a record). With these foundations in place, the next chapter reviews the literature on GDPR compliance models and identifies the questions that remain uncovered.

Chapter 3

Literature Review

This chapter presents the analysis of the current approaches in the literature addressing the tasks involved in representing, querying, or checking compliance with the GDPR provisions. It meets the necessity introduced by the second step of the used methodology, which recommends reusing existing models. Therefore, we are interested in identifying and describing the approaches that can provide a vocabulary to help answer those competency questions elicited in Chapter 2. The relevance of this chapter also lies in the imperative necessity to understand the comparable approaches to analyze them against the competency questions: does the current literature provide a sufficient vocabulary to answer this competency question? What concept can be reused, and what must be proposed? Targeting these questions and guided by a systematic methodology for the literature review, we collect a non-exhaustive set of approaches addressing the GDPR provisions and an analysis comparing approaches obtained against each competency question presented in Chapter 2.

3.1 Methodology

In the second step of the methodology proposed by Noy and McGuinness (78), the authors encourage investigating and reutilizing the existing contributions in the literature. This step thus leads to conducting a systematic mapping of the literature.

The primary objective of systematic mapping, according to Petersen et al. (98), is inter alias to capture a snapshot of a research area. Within this systematic mapping, our expectations are twofold: identify the current contributions that address the GDPR compliance issues proposing models to represent the GDPR provisions, particularly utilizing semantic technologies.

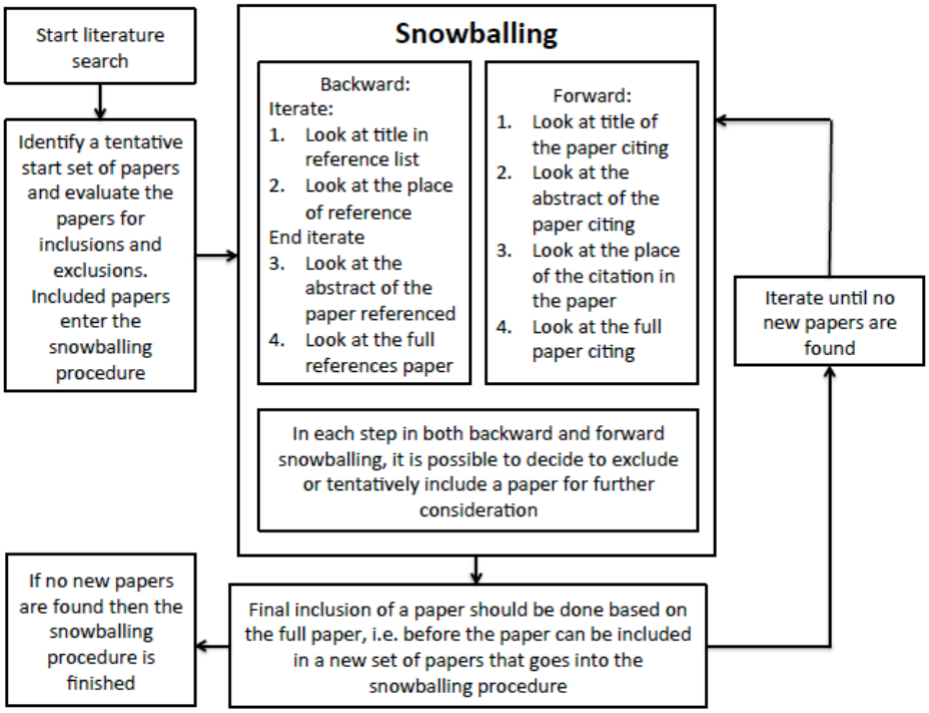

Figure 3.1: Snowballing steps (published by Wohlin(113)).

The performance of this mapping leans on the Wohlin (113) methodology. He has proposed a method to conduct a systematic mapping of the literature, namely Snowballing. Snowballing initializes with a given start set of papers related to the objectives and a set of inclusion or exclusion criteria that determine the incorporation or rejection of further articles into this set of related papers. Once it is defined the start set of papers and the inclusion/exclusion criteria, the second step consists of "rolling" this start set one step backward (i.e., individually analyzing each document referenced by each paper in our starter set) or one step forward (i.e., analyzing each article that cites those in our starter set). Each analysis activity consists of checking whether the document in hand meets the defined inclusion/exclusion criteria. If the document meets the inclusion criteria (or if it does not meet the exclusion criteria), it is put together with the set of related papers. Otherwise, the document is discarded. At the end of this process, it expects we hold a new set of associated documents, which can be submitted to this previous step as much as this current step reveals new non-analyzed documents. Figure 3.1 illustrates all Snowballing methodology steps.

Both Wohlin(113) and Petersen et al.(98) have suggested starting with a large and diverse set of papers (regarding authors, conferences, publishers, etc.) as possible, due to the method to discover new documents, which depends on the previous already reached. The outcomes quality of this methodology thus is higher as diverse as this initial set of papers. To achieve such diversity, we have compounded this first start set by conducting a tiny system-

atic mapping by searching the research term drawn ahead on Google Scholar.

Research Term (RT): `("General Data Protection Regulation" OR GDPR) compliance ("semantic technology" OR ontology OR provenance)`

This term led us to analyze the first three pages of Google Scholar until page three, where new readings produced no further results. Within this effort, we checked a total of 30 papers, in which only six(48; 83; 84; 112; 28; 86) were selected according to the inclusion criteria, namely:

1. The approach should propose a data provenance model to represent activities information related to GDPR obligations;
2. The proposed model should be publicly available so that we can compare it against the other approaches;
3. The document should be written in English or Portuguese;
4. The document should be publicly available or at least accessible under those digital libraries CAPES CAFE has signed;
5. It should be a peer-reviewed document (e.g., we have discarded arXiv as a source);

Starting from this initial set of documents(48; 83; 84; 112; 28; 86), we start our systematic mapping by rolling one step backward. By performing this task, we analyze a total of 178 documents, through which eight(16; 110; 22; 109; 4; 95; 94; 65) match within the elicited inclusion criteria. From these 14 papers thus, we conduct one step forward, seeking articles that cite these. This effort led us to analyze a total of 112 documents, through which four(71; 23; 111; 106) match within the inclusion criteria.

Overall, we have analyzed a total of 353 documents; 34 of which are not made publicly accessible or the model proposed is not published, 16 are discarded due to not clearly presenting whether they were submitted to a peer-reviewed process, and 303 remaining documents laid on the first criteria, present a different objective than that we are seeking with this work.

Any further information about this step is further explained in Chapter 3. We have conducted this study leaning on the Snowballing[1] project, which contains all information about this mapping.

[1]The Snowballing is available at https://github.com/JoaoFelipe/snowballing. All metadata about our mapping was made publicly available in `https://github.com/danielpcampagna/master-the-state-of-the-art`

3.2 Approaches

In this section, we summarize the publications we have found in the literature that match with previously listed criteria. Although GDPR was published only in 2016, from January 2012, the European Commission released the official proposal for this regulation, marking the formal start of the legislative process. Therefore, the oldest publications date prior to the date the GDPR was published. We present the results of this survey in a logical-temporal way, starting from publications in 2015 until 2024.

In 2015, (16) sowed the first seeds on the use of semantic models to address the data protection requirements derived from the GDPR, while the Regulation was still in draft form. They propose a bottom-up ontology that organizes the domain into three blocks (basic data protection principles, rules of processing, and data-subject rights) and articulates the deontic relations between the data subjects' rights and the corresponding duties of controllers and processors. The ontology is developed under the METHONTOLOGY and NeOn specification tasks: requirements are elicited as seven informal competency questions defined by the authors (such as "What are the obligations of a data controller?"), the conceptualization reuses concepts from LKIF Core and the Simple Knowledge Organization System (SKOS)[2], the implementation uses OWL on Protégé, and the evaluation pairs the OOPS! pitfall scanner with SPARQL queries that translate those competency questions. To demonstrate the practical value of the model, the authors extend Business Process Model and Notation (BPMN) with a new *Data Protection Task* type and an Eclipse plugin that injects the ontology's concepts as workflow annotations, enabling a Privacy-by-Design style of integration. Because the publication predates the final version of the Regulation, the authors explicitly characterize the ontology as a work in progress that requires further refinement once the legislative text is finalized. Methodologically, the work falls primarily into the *CQ-driven (author-defined)* group, even though the legal knowledge is harvested from normative frameworks and aligned with reused vocabularies.

Two years later, during the transition period before the GDPR became applicable, (48) proposed the Consent and Data Management Model (CDMM), a semantic model designed to capture information regarding consent, such as the context in which it was granted or revoked, the associated permissions, and the activities involved in fulfilling obligations under that consent. The model innovates by extending the PROV-O ontology (**?**) to represent consent provenance and by sharing the ontology through Semantic Web Technologies. The authors define the intertwined lifecycles of consent and data, illustrating how data process-

[2] `https://www.w3.org/2004/02/skos/`

ing should occur only after consent is verified, and demonstrate the model's applicability through a use case scenario. However, the validation remains a preliminary proof of concept, and there is no explicit path from the GDPR requirements to the proposed elements, a limitation the authors themselves acknowledge by characterizing the model as a generic upper ontology.

In 2018, Pandit et al. made two interconnected contributions. The first (94) addresses the difficulty of understanding privacy policies, which are legal documents describing activities over personal data such as collection, usage, processing, sharing, and storage. The approach leverages keyword-based entity extraction using GDPR terms and concepts, representing the extracted data via the GDPRov ontology (92). This extraction is fundamentally supported by their second contribution, GDPRtEXT (89), a comprehensive linked data resource that models the GDPR text using the European Legislation Identifier (ELI) (**?**) and provides a SKOS-based vocabulary defining over 200 terms and concepts.

The second main contribution of (91; 95; 96; 85) targets the detection and representation of changes in the context of consent and activities for GDPR compliance. This contribution centers on GDPRov, a data provenance model designed to represent metadata regarding data changes. It innovates by capturing both planned (ex-ante, prospective) and executed (ex-post, retrospective) activities to demonstrate GDPR compliance. The authors extend GDPRov with concepts from P-Plan[3] (an extension of PROV-O) for representing the provenance of activities, and the Open Digital Rights Language (ODRL)[4] for representing consent. Although GDPRov was developed to address some official competency questions present in the Guidance for SMEs (36), its final form left open almost half of these questions, which motivates the present work.

In 2018, (112) advanced prior ontological efforts on GDPR modeling, such as that of (16), by extending the W3C PROV-DM (**?**) to represent GDPR concepts and incorporating temporal reasoning so that their model can track provenance information of activities, consent, and personal data under GDPR. The authors also propose design patterns to assist model adoption. Further, in 2020, by applying a non-systematic interpretation of GDPR, (28) identify 11 limitations in Ujcich's model, for which they propose extensions to fulfill the original gaps and provide two design patterns as a demonstration of their model's practical appli-

[3] `https://vocab.linkeddata.es/p-plan/index.html`

[4] ODRL is a W3C Recommendation defined by three companion resources: the Core Information Model (`https://www.w3.org/TR/odrl-model/`), which specifies the abstract syntax and semantics; the Vocabulary and Expression specification (`https://www.w3.org/TR/odrl-vocab/`), which enumerates the standard terms; and the ontology namespace (`https://www.w3.org/ns/odrl/2/ODRL20.html`), which publishes the machine-readable RDF vocabulary.

cability. A shared limitation across these approaches is the reliance on the authors' non-systematic interpretation of the GDPR to derive the requirements and limitations, rather than a systematic or official methodology rooted in the legal text itself.

Alongside these, the SPECIAL H2020 project[5] concretized Semantic Web standards into a scalable compliance architecture. In 2018, (65) published the SPECIAL consent, transparency and compliance system, introducing two privacy vocabularies[6] that model the collected data, privacy policies, consent management, purposes, and data processing activities. Building on this architecture, in 2020 (23) formalized the SPECIAL policy language as a tractable fragment of OWL2 (termed Policy Logic), enabling the formal representation of data subject consent, business policies, and regulatory obligations. The language was derived from a detailed analysis of the GDPR text to identify deontic concepts such as prohibitions, permissions, and obligations, and to map six core data usage categories (purpose, data categories, recipients, storage duration, etc.) based on GDPR Article 30. To ensure interoperability, the authors adapted existing standards such as P3P and ODRL, and collaborated with the W3C Data Privacy Vocabularies and Controls Community Group (DPVCG). Compliance is checked through structural subsumption, and a specialized reasoning algorithm (PLR) achieves the scalability necessary for real-time verification, evaluated on real-world pilots handling millions of compliance checks per day. Although these works pioneered the use of Semantic Web standards in a scalable architecture, the publications do not detail the engineering process behind the vocabularies, which makes it difficult to assess whether they answer the open competency questions that motivate our work.

Also in 2018, (110) introduce a conceptual model providing a concise visual representation of the GDPR's entities, actions, artifacts, and rights using Unified Modeling Language (UML) notation. The model provides a comprehensive overview by mapping classes (such as Consent, PersonalData categories including sensitive data, and TechnicalMeasures) directly to specific articles of the regulation. The authors demonstrate that the model can serve as a foundation for organizational privacy management, aiding in the extraction of compliance rules and the identification of regulatory gaps in data processing activities. While the model is intended to drive extensions in modeling languages like BPMN, the authors acknowledge its preliminary nature, noting the need for formalized constraints and further validation from legal experts.

[5]SPECIAL (Scalable Policy-aware Linked Data Architecture For Privacy, Transparency and Compliance) is a European Horizon 2020 project focused on developing technical solutions to address data protection challenges in big data use cases. See more in `https://www.specialprivacy.eu/`

[6]the SPECIAL usage policy language is publicly available at `http://purl.org/specialprivacy/policylanguage`, while the log vocabulary is available at `http://purl.org/specialprivacy/splog`.

In 2019, (19) introduced the Privacy-aware Clinical Workflow (PaCW) ontology to facilitate automated GDPR compliance checking within the healthcare domain. The framework utilizes an ontology stack bridging three distinct domains: a Privacy Ontology, a BPMN Ontology for business processes, and a Clinical Domain Ontology. The Privacy Ontology formalizes four core GDPR principles: purpose specification, data minimization, consent check, and limited retention periods. This approach interprets the GDPR and aligns its duties with the existing BPMN 2.0 standard to enable compliance verification. Similarly, our work adopts an ontology-stacking strategy (combining GDPRov, the Trinity College Dublin ontology, and SAPOS) to cover the competency questions we address. Most of the components we reuse were already proposed in GDPRov.

Also in 2019, (40) proposed a semantic model for Data Flow Diagrams (DFDs)[7] designed to bridge the gap between software development and compliance with external policies such as the GDPR. The ontology, implemented in OWL 2, introduces formal predicates to represent data flows, processes, data stores, and interfaces, enabling the transformation of traditional design artifacts into machine-readable RDF representations. To illustrate the model's potential, the authors discuss how existing vocabularies such as GDPRov (92) and Ontology Design Patterns (ODPs) for personal data could be used alongside the DFD to enable SPARQL-based compliance queries. The ontology was released publicly under a CC-BY-4.0 license to support further integration into data governance frameworks.

In 2021, (111) proposed a comprehensive conceptual model for the GDPR using Model-Driven Engineering to support automated data protection and privacy compliance. The model is structured into nine specialized packages: Principles, Data Processing, Data Subjects, Main Actors, Data Subject Rights, Administration, Compliance, Data Transfer, and Enumerations. This framework covers a wide regulatory scope, with classes explicitly traced to GDPR articles spanning Chapters I through VIII. To facilitate automated scrutiny, the authors extracted 35 compliance rules from the legal text, providing them in both plain English and formal Object Constraint Language (OCL) constraints. Additionally, the model accounts for legal variability by defining 20 points where Member State laws may diverge, offering OCL-based resolutions to handle these jurisdictional differences. Although Torre et al.'s compliance-rule catalogue could inform our work, its conceptual-modeling focus on design-time verification diverges from our provenance-based focus on execution-time observability.

Also in 2021, (106) demonstrate the feasibility of using semantic technologies to con-

[7] Publicly available at `https://w3id.org/dfd`

struct a Register of Processing Activities (ROPA) from heterogeneous data sources. Beyond being mandatory under Article 30 of the GDPR, ROPAs are essential for Data Protection Officer (DPO) investigations and regulatory compliance monitoring. Their proposed framework, DPCat, is a profile of the DCAT-AP standard (37) that incorporates terms from the Data Privacy Vocabulary (DPV) (**?** 88) to represent processing purposes, data categories, and recipient locations at the metadata level. This approach is closely aligned with the one we adopt in this work.

In 2022, (53) addressed the systemic flaws in the traditional notice-and-consent model, arguing that current privacy mechanisms fail to empower data subjects to proactively manage their personal information. To address these limitations, the paper introduces a model for informed consent and an architecture for usable informational self-determination, supported by a semantic ontology designed to enable automated matching of user-defined Personal Privacy Policies (PPPo) against provider notices. The ontology was developed through a systematic five-step process: scope and objective identification, knowledge acquisition via an extensive survey of existing ontologies (CDMM, PrOnto (84), GConsent (87)) to identify gaps, UML-based conceptualization, implementation in Protégé, and a planned validation using Competency Questions. Methodologically, this work follows an interpretation-based approach, building a stand-alone ontology by interpreting the GDPR requirements for informed consent and aligning them with the proposed architecture.

In 2022, (68) introduces a knowledge graph-based approach to model and manage informed consent through its entire life-cycle, addressing deficiencies in transparency and human-machine interpretability. This research establishes the smashHitCore ontology, which extends the GConsent model by incorporating consent provenance, storage status, and granular details regarding sensor data and data processing activities. Developed for cross-domain data sharing in smart city and insurance contexts, the model was validated using competency questions derived from specific GDPR articles in collaboration with legal and industry experts. While Kurteva's work focuses on representing the consent life-cycle and its comprehension by humans and machines, GDPRov targets the detection and representation of changes in consent and processing activities for provenance purposes; together, both models offer complementary views on consent management.

(69) address the challenge of translating legislative requirements into software practices by proposing PRIAM (PRIvacy Assessment Model), a user-centric framework designed to assist Privacy-by-Design (PbD) and ensure GDPR compliance from the initial design stages. Instead of using PROV-O to detect and represent the changes in consent and processing

across planned (ex-ante) and executed (ex-post) phases, PRIAM focuses on embedding these compliance requirements into the software architecture from the outset. The authors characterize PRIAM as a preliminary description lacking detailed engineering evaluation, which limits comparability with models such as GDPRov that have been validated against official competency questions.

In 2022, Negri-Ribalta et al. (76) address the complexities of GDPR compliance by extending the Socio-Technical Security modeling language (STS-ml) (39; 82). STS-ml had been previously adapted by Robol et al. (102) to identify personal data, data controllers, processors, and legal bases. Using the single case mechanism experiment technique, the authors exposed the STS-ml metamodel to privacy cases representing each of the seven GDPR principles. They identified gaps where the language could not capture critical compliance information and proposed five lightweight attributes: retention time, special categories of personal data, asymmetrical relationships, provider of minor information, and EU actor. These attributes are grounded in the GDPR itself, data protection authority interpretations, and ontologies such as PrOnto. Among them, retention time directly addresses our competency questions CQ8, CQ28, CQ29, and CQ51 by attaching a monthly integer to the legal basis class. However, the representation granularity differs from our contribution, where retention constraints require richer temporal constructs and execution-time observability.

In 2022, Ryan and Brennan (104) propose the Common Semantic Model for ROPA (CSM-ROPA), a machine-readable mediation layer designed to improve GDPR accountability and streamline ROPA maintenance. Developed using Action Design Research (**?**) and built as a profile of the DPV, CSM-ROPA addresses the heterogeneity of data sources and lack of interoperability between organisations and regulators. An evaluation against the UK Information Commissioner's Office (ICO) accountability framework demonstrated 98% expressivity of the 139 unique terms identified in official compliance requirements. While CSM-ROPA provides broad semantic coverage of what a ROPA should record, tracking whether those obligations were actually fulfilled at execution time requires a provenance perspective that our contribution addresses.

In 2022, Chhetri et al. (31) propose an automated compliance checking tool that translates GDPR requirements into software through a regulation-to-code process: legal text is first mapped to Standard Data Protection Model protection goals, then to Technical and Organizational Measures, and finally to code in collaboration with legal experts[8]. The tool architecture follows a microservices pattern with serverless functions for logging and query-

[8]The tool's source code is publicly available at `https://github.com/tekrajchhetri/GDPR_compliance_tool`

ing, a semantic-preserving storage layer built upon SPARQL-based data management, and the smashHitCore ontology[9] to represent informed consent, processing activities, and involved entities. Key functionalities include consent lifecycle management, periodic automated compliance checks, and auditing. The approach demonstrates the feasibility of concretizing semantic models into operational tools for compliance assessment, bridging the gap between conceptual ontologies and practical deployment.

Also in 2022, by employing term extraction techniques, semantic analysis, term frequency enumeration, de-duplication, and antonym/homonym identification on 17 guidance documents aimed to assist organizations with maintaining their ROPA documents, provided by Data Protection Authorities[10], Ryan et al. (105; 104) yielded CSM-ROPA. CSM-ROPA was designed to provide an interoperable machine-readable vocabulary to represent all DPA-specified ROPA information. The proposed model comprises 47 unique concepts, of which 44 match exactly with terms in DPV, one partially matches (i.e., **dpv:PersonalDataHandling**) and two are new terms introduced by the authors into the DPV (i.e., **dpv:Status** and **dpcat:DataBreachRecord**)[11]. Further, by analyzing five use cases that describe the data flows between stakeholders, the authors identified additional requirements to enable the representation of a ROPA in practice, and proposed the data processing catalog (DPCat [12]), which extends Data Catalog Vocabulary (DCAT)[13] (37; 11) and CSM-ROPA for representing catalogs and datasets information. DPCat introduced 12 classes and 48 relations. However, the entire CSM-ROPA and DPCat models have already been absorbed in the second version of the DPV. While the approach provides a systematic method for harmonising heterogeneous DPA templates into interoperable compliance records, it does not incorporate the GDPR Readiness Checklist Tools (36) provided by the Irish Data Protection Commission, as Ireland was not among the jurisdictions whose templates were analysed.

In 2023, Gupta and Hahmann (60) propose the Ontology for Privacy Policies of OSNs (OPPO)[14] to formalise the granular data practices outlined in Online Social Network privacy policies. Grounded in the Basic Formal Ontology (BFO) and its extensions, the ontology introduces 60 classes and relations developed following the METHONTOLOGY framework and

[9] Publicly available at `https://smashhiteu.github.io/smashHitCore/`

[10] The GDPR establishes a total of 31 Data Protection Authorities (DPAs) representing nations and member states from the EU. According to the authors, only 17 of 31 DPAs have published guidance documents (ROPA templates). The 17 EU members from which these DPAs belongs are the following: Belgium, Greece, United Kingdom, Poland, Cyprus, France, Portugal, Germany, Denmark, Luxembourg, Finland, Czechia, Croatia, Italy, Lithuania, Slovenia, Slovakia

[11] Both new terms have become part of DPV since version 2.0

[12] Publicly available at `https://w3id.org/dpcat`

[13] a W3C standard for facilitating interoperability between data catalogues, made publicly available at `https://www.w3.org/TR/vocab-dcat-2/`

[14] Publicly available at `https://github.com/SanondaDattaGupta/OPPO-Ontology`

guided by 45 author-defined competency questions derived from a manual analysis of the GDPR, CCPA, and the privacy policies of ten major OSNs. OPPO enables automated reasoning and querying via SPARQL and the HermiT reasoner, demonstrated through a proof-of-concept with Telegram's privacy policy. While the ontology models data storage, retention, and security practices, its scope is limited to policy content rather than tracking provenance of actual processing activities, making it complementary to our approach.

In 2023, Echenim and Joshi (43) introduce the IoT-Reg Ontology, a semantic framework to automate privacy compliance for wearable IoT devices by unifying requirements from the GDPR, HIPAA, and NISTIR 8228[15]. The model adopts a stratified approach aligned with the IoT data lifecycle (collection, retention, processing, sharing, and deletion) and incorporates components from established ontologies, including the Semantic Sensor Network (SSN) and Sensor, Observation, Sample, and Actuator (SOSA) ontologies, to represent sensing activities and device capabilities. The ontology was developed following the Noy and McGuinness methodology (78) and validated through deterministic competency questions, Description Logic rules, Semantic Web Rule Language (SWRL) rules, and the Pellet reasoner in Protégé. Although it addresses retention periods, data accuracy procedures, and systematic erasure through its lifecycle stages, the ontology is not publicly available, which limits direct comparison with our model.

In 2023, Ruiz et al. (103) bridge the gap between the legal requirements of GDPR Article 35 and technical implementation guidelines by mapping Data Protection Impact Assessment (DPIA) concepts into established assurance frameworks[16] to facilitate mutual comprehension between legal experts and system engineers. This systematic mapping was performed using the Common Assurance and Certification Metamodel (CACM) (8) for modeling reference frameworks and a qualitative content analysis methodology(72) to manually yet systematically extract and annotate concepts from the legal text. The resulting models were iteratively refined by a multidisciplinary focus group of privacy experts to ensure their soundness for privacy-related system assurance.

In 2024, Pandit et al. (88; 64) released DPV version 2.0, evolving the previous version. DPV was originally designed to provide an interoperable vocabulary for representing personal data and its processing under the GDPR. The major changes in DPV 2.0 expand the scope from personal data to any data or technology, including Artificial Intelligence. The same semantic structure now covers both personal and non-personal data across regula-

[15] NISTIR 8228 is a U.S. National Institute of Standards and Technology publication that establishes high-level goals for protecting device security, data security, and individuals' privacy throughout the IoT lifecycle (21)

[16] Specifically, the authors approach the ISO/IEC 29134 (Information technology—Security techniques—Guidelines for privacy impact assessment) and the EU Smart Grid DPIA template.

tions such as the GDPR, the Data Governance Act (DGA), and the AI Act. The update removed 56 concepts to dedicated extensions and introduced 311 new ones (911 total in the core DPV), with redesigned namespaces to distinguish jurisdictions and laws. The evolution was managed by the W3C DPVCG, where members propose additions or modifications via GitHub issues, followed by discussions, expert advice, and community voting to reach consensus. While DPV provides a consensus-driven vocabulary for privacy concepts, it is not a provenance model and therefore cannot determine whether processing obligations were actually fulfilled at execution time. A point-by-point comparison against the DPV 2.0 changelog confirms that our contribution remains largely complementary. The duration and storage-condition terms our extension relies on (`:TemporalDuration`, `:UntilTimeDuration`, `:UntilEventDuration`, `:EndlessDuration`, and `:StorageCondition`) are reused from DPV rather than authored here, and DPV 2.0 neither revises nor supersedes them, so the alignment we build upon remains valid. What DPV does not offer, in either version, is a way to attach those terms to the plans and executions of a provenance model, and supplying that bridge is what our extension contributes on this front. Conversely, DPV 2.0 does introduce new concepts for data deletion and erasure policies (`dpv:Delete`, `dpv:DataDeletionPolicy`, `dpv:DataErasurePolicy`) and for data-quality management (`dpv:DataQualityAssessment`, `dpv:CorrectingProcess`), which intersect respectively with our CQ51 and CQ28 contributions and could be aligned in future work. Finally, our `:FixedOccurencesDuration` corresponds to the existing `dpv:FixedOccurrencesDuration`, whose original spelling was corrected in DPV 2.0.

Looking across the works reviewed above, we can cluster them by the methodology each author used to derive their semantic model into five groups. A first group is driven by competency questions and itself splits in two flavours: those anchored in *official check-lists* (91; 95; 96; 85), which deliver greater traceability to regulatory expectations, and those relying on *questions defined by the authors* (16; 68; 60; 43), which offer greater flexibility but weaker traceability. A *directive-driven* (104) angle, in which terminology is extracted directly from official directives that do not explicitly pose questions. Every approach that begins from official text also performs interpretive alignment, and so lands in one of the groups below. A *community-driven* group centres on standardization efforts, most prominently DPV (**?** 88), whose strength is consensus and broad reuse but whose traceability of each component back to a single regulatory term is harder to establish. Finally, the largest group adopts an *interpretation-based* methodology, building stand-alone ontologies or aligning GDPR duties to existing models (48; 94; 89; 112; 40; 28; 65; 23; 110; 19; 111; 106; 53; 69; 76; 105; 31; 103), which maximises flexibility at the cost of reproducibility.

Beyond traceability and flexibility, these groups also differ along three practical axes that bear directly on a controller's ability to adopt and sustain a model: *modeling effort* (interpretation-based approaches demand the most up-front analysis, whereas CQ-driven approaches front-load that effort and amortize it across instantiations, and community-driven reuse is cheapest to start from but defers effort to alignment); *maintainability* (community-driven vocabularies inherit the upkeep of an active standardization body, while interpretation-based and author-defined CQ models tend to stall after publication); and *applicability to real scenarios* (directive- and official-checklist CQ-driven models map onto instruments controllers already use for self-assessment, while interpretation-based models are more expressive but harder to audit against an external reference). No group dominates on every axis. In this dissertation we adopt an official-checklist CQ-driven approach as the backbone because auditability against an external reference and reuse of the elicitation effort matter most for our goal of answering officially posed compliance questions, complemented by community-driven vocabularies that contribute consensus and low maintenance; we acknowledge that this choice inherits the limitations of its source instrument, namely that the coverage is bounded by the questions the checklist authors thought to ask and that a checklist may lag behind regulatory or jurisprudential change.

3.3 Answering Our RQ

Having surveyed the state of the art, we can now return to the research question (**RQ**) posed in Chapter 1: to what extent can data provenance models answer the official GDPR compliance questions that current models leave uncovered? The works reviewed above already speak to it. Data provenance has repeatedly been employed to represent GDPR concepts and to support compliance: Fatema et al. (48) extend PROV-O to capture the provenance of consent; Ujcich et al. (112) extend PROV-DM with temporal reasoning to track activities, consent, and personal data; and Pandit (85) proposes GDPRov, which combines P-Plan and ODRL to represent both the prospective and the retrospective provenance of processing. The feasibility premise underlying our RQ is therefore not in question: provenance can, in principle, express the information needed to check compliance.

What the state of the art does *not* yet provide is a complete answer against an official, externally defined reference. Most models derive their requirements from the authors' own interpretation of the legal text rather than from an official instrument, which makes their coverage difficult to audit against a common benchmark. Only a handful of works ground their requirements in official sources, as we do, and even these leave the questions we ad-

dress open. The closest is GDPRov itself (85): it is evaluated against the very checklist we adopt, the GDPR Guidance for SMEs (36), yet in its published form it answers only 32 of that checklist's 65 questions, leaving the other 33 open, among them the four we take up (CQ8, CQ28, CQ29, and CQ51). Ryan et al. (104; 105) likewise build on official material, deriving CSM-ROPA and DPCat from the records-of-processing templates published by 17 European data protection authorities; their aim, however, is to harmonize records of processing rather than to answer compliance questions, and their sources exclude the Irish DPC checklist, so our four questions lie outside their scope. In short, provenance *can* be employed, but even the works that share our official, checklist-driven footing stop short of answering the questions we address, leaving open just how far provenance can go against the compliance questions that supervisory authorities actually publish.

This is precisely the opening our work targets. Rather than building a new model, we adopt an official-checklist, CQ-driven approach and extend the most complete existing provenance model to answer specific official questions it currently leaves open. The extension that delivers this answer is the subject of Chapter 4.

3.4 Chapter Summary

This chapter surveyed the literature on models that represent GDPR concepts to support compliance, following the systematic method described in Section 3.1. We organized the reviewed approaches around the methodologies they employ and, crucially, around the provenance of their competency questions. A recurring finding is that most models derive their requirements from the authors' own reading of the GDPR rather than from an official instrument, which makes their coverage hard to audit against a shared benchmark. A few works do ground their requirements in official sources, and among them GDPRov (85) is the most relevant to us: it is evaluated against the same GDPR Guidance for SMEs checklist we adopt, but answers only 32 of its 65 questions. That gap, and the opportunity it represents, is what motivates the extension we develop next. In Chapter 4, we take GDPRov as our starting point and address four of its uncovered questions, reusing established vocabularies and adding the causal representation that links purpose fulfilment to data deletion.

Chapter 4

An extension to the GDPRov

4.1 Methodology

This research introduces an extended version of GDPRov, the provenance ontology proposed by Pandit (85), designed to address additional compliance questions relevant to SMEs (36) that previous versions have not covered. This result is obtained by employing traditional and well-established methodologies in ontology engineering (107; 78). The roadmap to achieve the objective of this study was based on a combination of well-established knowledge-engineering methodologies (78; 107). The first one, proposed by Noy and McGuinness (78), suggests a seven-step guideline to develop an ontology for a specific domain from scratch. They have defined ontology as documentation in a structured, clear, and unambiguous manner of concepts described by their properties and constraints that delimit each concept in a field or subject (78). For the purpose of this dissertation, we have assumed "ontology" and "model" as interchangeable terms [1].

The second methodology, introduced by Suárez et al. (NeON) (107), goes one step forward in ontology-engineering scenarios by identifying and suggesting systematic measures where reusing is necessary. Unlike the former, Suárez et al. are more concerned about those scenarios beyond the development from scratch. By this motivation, they have proposed non-rigid guidelines addressing nine strategies for such construction. Since our objective also includes the extension of prior models, the NeON methodology has also been shown as useful to this study as the first one.

[1] Noy and McGuinness (78), even following Gruber (59) definition of ontology, distinguish both terms. Atkinson et al. (13), analyzing the distinction between model and ontology (including Gruber's definition and Noy and McGuinness' distinction), have demonstrated that both terms can not be distinguished according to the literature. Based on Atkinson et al., we assume that ontology and model, in our case, are interchangeable terms. We further extend this explanation in Section 2.

Slightly adapting these methodologies according to our objectives (e.g., compressing and skipping some unuseful steps for our case), this study has been guided by the following four steps. Each one is detailed in its respective chapter.

The four steps consist of the following:

1. **Define the model scope.** Noy and McGuinness's methodology (78) starts by suggesting the task of elucidating the model boundaries as clearly as possible: what it needs to represent, for what purpose?

   As mentioned in Section 1.2.3, we have defined the model scope tight to those questions elicited in the GDPR Guidance for SMEs (36) that the current literature does not cover yet. Indeed, the methodology (78) indicates the adoption of competency questions as a means to determine the domain scope.

2. **Reuse existing models.** The second step of Noy and McGuinness's methodology (78) is to explore existing approaches. Since every conceptual model aims to be reusable, it is mandatory to assert whether such domains have not been covered yet by other similar domains. Indeed, both methodologies prescribe it. What parts of this domain have already been represented? What concepts lack representation?

   For the scope of this study, we have conducted a small systematic mapping of the literature about the related topic. Following a well-established methodology (113) of searching the existing model, we analyzed a total of 353 documents, from where eighteen papers were found related to the objectives of this study. Each of them was compared against the compliance questions. As a result, we verified that there still exists uncovered questions.

3. **Identify and describe the key concepts of the domain.** Once we held reusable concepts and identified lacks in the current literature, by the suggestion of both methodologies, we have introduced new concepts for this domain. This comprehends steps 2-6 of Noy and McGuinness's method (78). NeON (107), though, details different tips for reusing and merging tasks. In this step, we seek to characterize the lacking terms that should be explained to the model user: how is this concept used in this domain? How could they be hierarchically related? How can different sources of knowledge be merged?

   Following a systematic workflow, we have proposed new concepts that should be employed in any model to represent those targeted questions. Since we have detailed the entire process, starting from the competency question to their necessary concepts, we

consider these outcomes useful for future models aiming to represent these questions. At least, we think these outcomes can be quickly criticized.

4. **Evaluate the model.** Once we have gotten a version of the model, we go back to the scope initially elicited and check whether the obtained components are capable of representing that. Noy and McGuinness propose the use of instances to represent use cases. NeON goes one step further; in addition to evaluating whether we are producing the correct ontology, he also suggests assessing whether the process to get the model has been followed correctly: Beyond the capability of answering the competency questions by utilizing the proposed concepts, one must assess the quality of the model's design of the concepts and relationships. This last evaluation aims at analyzing modeling errors, such as polysemous elements or synonyms as concepts.

   The evaluation of the model's suitability is taken twofold: by instantiating some use cases run in prior work into the proposed model and then by representing the competency questions as SPARQL queries. Conversely, the assessment of the model's quality is taken by utilizing the OOPS! [2] (100) (OntOlogy Pitfall Scanner!), an online and methodology-independent ontology development tool that was built after its authors analyzed over 693 works. By considering a given ontology, OOPS! responds with a list of pitfalls it has detected in this ontology, classified into different importance levels. It is worth noting that OOPS! covers design and modeling pitfalls only, since it checks an ontology against a catalog of structural and semantic anti-patterns rather than performing logical reasoning; logical inconsistencies would require a description-logic reasoner instead.

Notwithstanding the construction of a model guided by systematic and well-established methodologies, it is fundamental to point out some principles tacitly present in every task of ontology engineering, according to Noy and McGuinness (78):

1. **There is no single correct way to model a domain.** The best solutions are often determined by the model application in the real world, which makes it inescapable to meet other suitable solutions.

2. **Ontology development is necessarily an iterative process.** The process always starts by crafting a simplistic first sample of the ontology. It goes to increasing the complexity of the ontology by revisiting the initial goals leading to possible redefinitions of initial concepts and the need to include a new one.

[2] `https://oops.linkeddata.es/`

3. **Concepts in the ontology should be close to objects (physical or logical) and relationships in the domain of interest.** The modeling process consists most likely of discovering the nouns (things) or verbs (relationships) hidden in the real-world domain.

Subsequently, for each identified term, we prioritized the reuse and adaptation of both ontological and non-ontological resources, drawing from the academic literature, community standards, and official publications, in line with best practices for ontology development (107). These newly identified resources were then integrated into GDPRov by specifying the relevant classes, properties, and instances and aligning them with existing taxonomies (78).

In this study, we apply this methodology to the questions CQ8, CQ28, CQ29, and CQ51 to demonstrate the still-unexplored potential of provenance and semantic web technologies in addressing compliance questions elicited by the (36). Although we do not address in detail the remaining uncovered questions of Guidelines (36) in this study, we propose five solution categories for addressing them.

The extended GDPRov ontology incorporates new elements from distinct existing ontologies. This process allows us to create a semantic model that bridges the gaps between official regulatory requirements and technical implementation. This extension is validated through a series of SPARQL queries that demonstrate its capability to answer the target compliance questions, providing SMEs (36) with a practical framework to assess and document their GDPR compliance obligations in a machine-readable format.

Because this reuse-first stance shapes what is genuinely ours in the result, the tables that present the extension carry a *Nature* column recording, for each construct, how much of it we authored. We distinguish three cases. A *reused* construct is one whose meaning is already fixed by an external vocabulary and which we declare in the GDPRov namespace only so that the model can refer to it; its semantics come from the source, not from us. An *alignment* adds an axiom relating a class that already exists (either in GDPRov or introduced earlier in our own extension) to an external one, without introducing any new concept. An *original* construct is one for which no counterpart exists in the vocabularies we reviewed, and which we therefore define ourselves. The balance among the three is deliberate: NeOn prescribes reusing an existing resource whenever one fits (107), so we introduce a new concept only where no established vocabulary supplies it. Our own contribution consequently lies in two places, namely the small set of original constructs and the alignment axioms that make the reused vocabularies operate inside a provenance model, which is something none of DPV, Time, SSN, or GDPRov provides on its own. Section 5.5 quantifies this balance across the four questions.

In the next sessions, we analyze each question and propose extensions to the GDPRov to address them.

4.2 Modeling Cost and Reproducibility

Extending the model to a new competency question carries a non-trivial manual cost: each question must be read, assessed for provenance-suitability, and decomposed into its constituent terms before any ontological resource can be reused. We perform this decomposition through Qualitative Content Analysis (103; 72), manually coding the regulatory text into candidate model elements. This cost is, however, bounded and front-loaded rather than open-ended. The procedure we follow is a fixed, well-established ontology-engineering pipeline (78; 107) (assess suitability, decompose into terms, reuse ontological and non-ontological resources, integrate and align, and validate via SPARQL), so the same steps recur unchanged for every question and the elicitation effort is amortized across instantiations. The labor bottleneck is concentrated in a single step, the extraction of terms from legal prose, which is a natural candidate for partial automation by language models acting as assistive tools; we return to this prospect in Section 6.

Not all of the resulting model is equally reusable, and it is worth separating what every instantiation needs from what is specific to the cases treated here. The *essential* part, reusable across controllers and jurisdictions, consists of the constructs drawn from community vocabularies: the DPV and Time duration classes that express retention periods (CQ8), the data-comparison and discrepancy constructs derived from `dpv:Match` (CQ28), and the SSN observability constructs that link purpose fulfilment to deletion (CQ29). The *optional*, case-specific part consists of interpretive parameters that a different deployment may legitimately set otherwise: the operationalization of "without delay" as one month (CQ28), of "systematic" as automated and pre-arranged (CQ51), the single-source-of-truth simplification (CQ28), the reading of CQ29 as prospective, and the inverse-query reformulation of CQ51. The essential constructs can be instantiated mechanically; only the optional parameters require a modeling decision.

This separation also delimits the reproducibility risk. The interpretive parameters are not hard-coded into the ontology: each is documented and justified against official guidance (Article 12 for "without delay" (1, Art. 12), and the WP29 and EDPB opinions for "systematic" (12) and for the storage-limitation conditional (47)), and each is exposed as an explicit, replaceable value rather than embedded in the class structure. Consequently, a controller

in another jurisdiction can re-parameterize these values without re-modeling the ontology. The residual risk, that a chosen interpretation may not survive future jurisprudence, is inherent to any formalization of legal language; we discuss it, and its mitigation, in Section 6.

4.3 Competency Question CQ8

CQ8: For each category of personal data, list the period for which the data will be retained e.g. one month? one year? As a general rule, data must be retained for no longer than is necessary for the purpose for which it was collected in the first place.

The expected answer to this question is a list of pairs, each consisting of a *category of personal data* and the corresponding *period for which the data will be retained*. This information is typically included in the privacy policy and can be incorporated into the prospective provenance graph by the controller. Since addressing this question requires representing only these two elements, we proceed to analyze them in detail.

4.3.1 The Terminology in Question CQ8

The Personal Data Category. GDPRov (85) defines categories of personal data as subclasses of `:PersonalData`. With this approach, the responsibility for specifying which data categories should be represented is deferred until the model is instantiated. Therefore, no further extension is necessary.

The Retention Period. The GDPR lacks a definition for the retention period. However, the literature indicates that the most commonly used approach is to use the legal bases that legitimize data collection to determine the retention period.

(19; 20) have proposed the `RetentionPolicy`class in their ontology to represent the data retention period. However, they lack an example of how to use this. By analyzing the existing privacy policies, (80), along with (52) and (**?**), have identified four possible cases that describe the retention period of a piece of personal data, namely:

1. *Fixed period*: when the retention period is explicit (e.g., 30 days). According to the analysis of how long privacy policies store user data by (80), this case accounts for 40% of the occurrences.

2. *Fixed date*: when the validity time is indicated (e.g., December 31st, 2022);
3. *After purpose*: when the data is stored as long as it is needed to perform a requested process;
4. *Indefinite*: when the retention period is not explicit, corresponding to 7% of the total (80).

4.3.2 GDPRov Extension to address CQ8

Once we have identified the terms that must be included in the GDPRov ontology to address CQ8, we have searched for the corresponding classes and properties that closely represent these terms in the other ontologies. DPV[3] and Time[4] ontologies contain elements that are very close to what we need. Therefore, we have extended the GDPRov ontology with some classes and properties from the DPV and Time ontologies that relate to these elements. The two tables presented in the remainder of this subsection, Tables 4.1 and 4.2, list the classes and properties, respectively, that we have reused to extend the GDPRov ontology.

As Table 4.1 depicts, the term *Fixed Period* is encoded into the GDPRov through the **:TemporalDuration** class[5]. We propose this class as a subclass of **dpv:TemporalDuration** so that it can inherit all the characteristics of the latter. Also, it extends the **:Duration** class, which is an extended class from DPV ontology (i.e., **time:Duration**) and aimed to encode a generic duration term. Similarly to the *Fixed Period* term, the **:TemporalDuration** class is used to represent a continuous and clearly bounded time interval of time, such as *6 months* or *2 years*.

The next three classes presented in Table 4.1, **:UntilTimeDuration**, **:UntilEventDuration**, and **:EndlessDuration**, encode the terms *Fixed date*, *After purpose*, and *Indefinite*, respectively. The sixth class, **:FixedOccurencesDuration**, is equivalent to the class of the same name proposed by the DPV ontology.

Finally, following the standard established by the DPV ontology, the **:StorageCondition** class should be used as the bridge between the **:Process** and the **:Duration** classes, expressing the *Storage Condition* term mentioned in CQ8, for which some data was col-

[3] Publicly available at `https://w3id.org/dpv`
[4] Publicly available at `https://www.w3.org/TR/owl-time/`
[5] Classes with no namespace prefix belong to the GDPRov namespace, either because GDPRov already defined them or because our extension declares them there. Residing in that namespace does not by itself make a construct ours: the *Nature* column is what separates the constructs reused from external vocabularies, the alignments we add, and the ones we author.

Class	Description	Nature
:Duration	Measurable temporal extent expressed as a decimal number scaled by a temporal unit (equivalent to time:Duration).	Reused
:TemporalDuration	A temporal duration aimed to continuously cover a fixed and pre-conceivable length of time; e.g., *6 months* (equivalent to dpv:TemporalDuration and subclass of :Duration).	Reused
:UntilTimeDuration	A duration that is defined by a specific, predetermined end date or point in time; e.g., *2026-12-31* (equivalent to dpv:UntilTimeDuration and subclass of :Duration).	Reused
:UntilEventDuration	A duration that continues until the occurrence of a specified event or condition, such as when data is no longer needed for its original purpose. (equivalent to dpv:UntilEventDuration and subclass of :Duration).	Reused
:EndlessDuration	A duration characterized by its indeterminate or intentionally perpetual nature, signifying an absence of a defined endpoint (equivalent to a instance of dpv:EndlessDuration and an instance of :Duration).	Reused
:FixedOccurencesDuration	A duration defined by a fixed number of repeated occurrences of an event or activity, regardless of the actual temporal length; e.g. *3 times a year"* (equivalent to dpv:FixedOccurencesDuration and sub class of :Duration).	Reused
:StorageCondition	Conditions required or followed regarding storage of data (equivalent to dpv:StorageCondition).	Reused

Table 4.1: Extended classes for GDPRov to address the CQ8. Every class in this table is *reused*: it is declared in the GDPRov namespace so the model can refer to it, but its meaning is fixed by the source vocabulary. `dpv:` and `time:` are prefixes for the ontologies defined in `https://w3id.org/dpv` and `https://www.w3.org/TR/owl-time`, respectively.

lected. All of these classes are connected through :hasStorageCondition and :hasDuration properties as described by Table 4.2.

It is worth being explicit about where the contribution lies here, since none of the constructs in Tables 4.1 and 4.2 is original to this work: each is reused from DPV or Time, as the *Nature* column records. What our extension contributes for CQ8 is the integration itself, that is, attaching DPV's storage-condition and duration vocabulary to the GDPRov :Process class, so that a retention period ceases to be a stand-alone annotation and becomes a property of a plan whose execution the provenance graph can later be queried against. Neither vocabulary offered this connection on its own: DPV is not a provenance model, and GDPRov carried no vocabulary for durations. The same reused constructs are what allow CQ51 to be

Property Domain	Target	Description	Nature
`:hasStorageCondition` `:Process`	`:StorageCondition`	Indicates the information about storage condition of a given process (equivalent to `dpv:hasStorageCondition`)	Reused
`:hasDuration` `:StorageCondition`	`:Duration`	Indicates the information about duration of a given process (equivalent to `dpv:hasDuration`).	Reused

Table 4.2: Extended properties for GDPRov to address the CQ8. Both properties are *reused* from DPV. `dpv:` is a prefix for the ontology defined in `https://w3id.org/dpv`.

answered later (Section 4.6) without introducing any further element.

4.4 Competency Question CQ28

CQ28: Are procedures in place to ensure personal data is kept up to date and accurate, and where a correction is required, the necessary changes are made without delay?

4.4.1 The Terminology in Question CQ28

This question may suggest burdening the controller with the guarantee that the personal data is kept *up-to-date* and *accurate* through the use of procedures. Additionally, it must be performed *without delay*. Answering this question means addressing these terms: how do we certify that a correction is required, that the necessary changes were made, and that they were done without delay. Both the literature and the GDPR corpus lack a clear definition of these terms, so we must dive into a positive analysis, considering these terms.

Up-to-date and Accurate Data. As far as we have researched, the GDPR corpus (including specialized text) lacks a clear definition of the terms *up-to-date* and *accurate*. Looking at the Cambridge Dictionary, the concepts of *accurate* and *up-to-date* can be understood as *something* in *accordance or agreement* with the *latest, the newest, or the most recent standard or model*. Although both terms are grammatically classified as adjectives, we want to underscore that they function as the predicate in a triple *subject-predicate-object* sentence, connecting the subject (i.e., the *something*) to the object (i.e., *standard or model*).

Diving into the analysis of *accurate*, the dictionary uses the prepositional phrase *in ac-*

cordance with or in agreement with to define it. This prepositional phrase implies the existence of an object (i.e., the *standard or model*) that is used to evaluate whether the sentence subject (i.e., *something*) is in accordance or agreement with it. Putting it differently, when we say "A is accurate", we tacitly suppose a B (in general, an ideal version of A) in such a way that makes it equivalent to saying "A is somehow equal to B". It is not unusual to measure the quality of being accurate (i.e., the accuracy of something) through a quantifier (e.g., "A is 73% accurate when compared to B"). Hence, the use of *accurate* always implies the existence of a comparable object and qualifies the sentence subject by establishing some kind of proximity to that object.

Proceeding to the analysis of the *up-to-date* term, it not only supposes the existence of such an object (which holds a certain degree of proximity or an absence of differentiation to the sentence subject) but also introduces a single temporal dimension line connecting the sentence subject and object, although they may be positioned at different points of this timeline. This temporal dimension is based on the dictionary definition, where all used adjectives (i.e., *latest*, *newest*, and *most recent*) are temporal adjectives. Additionally, these temporal adjectives also imply the existence of measurable variations of the subject over time, so that it is possible to differentiate the subject at different instants along the time. For the sake of clarity, when it is said that "A is up to date", it tacitly evokes in our minds both a present-instant A and an immediate-past-instant A (A') so that A is somehow equal/equivalent to A'. Since *up-to-date* is a transitive predicate, saying "A is up to date" means "there is no variation in A since it was last updated". Therefore, this analysis concludes the mandatory presence of a third element (B or A', the sentence object) when we employ the terms *up-to-date* and *accurate*.

Referring back to the question CQ28, both predicates connect *the personal data* (the subject of the sentence) to some sort of object. But what exactly is this object? We understand that this object should be something real, something beyond the controller's boundaries. To clarify this analysis, we refer to the entity that holds the truth as the source of truth (SoT). The SoT is, in reality, the final entity that possesses accurate and up-to-date information about the personal data in the controller's possession. Therefore, since the controller possesses an information that derived from the SoT, when the SoT changes, the controller no longer has accurate, up-to-date personal data (e.g., once Alice moved to a new address, the address she registered at the Retail Shop becomes inaccurate and out-of-date). Hence, the SoT provides the final answer about whether the personal data is accurate and up-to-date. For the sake of simplicity, we assume only problems with a single source of truth (SSoT). In the context of

GDPR, only the data subject or a third party can assume the role of SSoT[6].

Driving this conceptual discussion towards a technical realization within the ontology, GDPRov already defines several classes that can capture these roles, namely `:PersonalData`, `:Data`, `:Controller`, `:DataSubject`, and `:ThirdParty`. In our usage, the SSoT role is fulfilled by instances of either `:DataSubject` or `:ThirdParty`.

Finally, the second conclusion we can draw from this analysis is that both predicates, in the context of this question, bear the same meaning: they establish a relation of similarity between the data possessed by the controller and the information corresponding to reality. The current version of GDPRov needs classes to represent this information. However, we need to analyze additional terms before proposing a solution.

Procedures in place to ensure accuracy. The question begins by asking whether "there are procedures in place to ensure personal data is kept up to date and accurate." At first glance, one may naively assume that this requirement is satisfied by only demonstrating the existence of activities aimed at rectifying personal data. In other words, if the controller has established procedures that update the personal data values to the newer corrected values, then such a controller would be considered compliant with this question. Although GDPRov already provides classes to draw these rectification activities[7] Fortunately, GDPRov provides `:DataCollectionStep` and `:DataCollectionActivity` classes to represent, respectively, the prospective and retrospective aspects of these data collection procedures. Properties `:collectsData` and `prov:used` are designed to specify the relationships between the data collection procedure and the personal data involved.

Where a correction is required. Finally, between the data collection and rectification activ-

[6] In cases where the controller act as the SSoT, the assessment is necessarily tautological, since the controller, who is seeking the truth, also acts as the source of truth. Thus, it is unnecessary to face this question.

[7] The `:RectifyData` class represents the prospective rectification process, `:RectifyDataActivity` the retrospective one, and `:DataRectificationProcess` is designed to represent the plan for rectification procedures. We want to shed light on the necessity of additional elements in GDPRov to address this question properly.

Because the rectification process requires collecting data from the SoT to evaluate the accuracy and currency of the current data, the mere existence of procedures for rectifying such data is insufficient. Hence, to properly address this question, it is also necessary to demonstrate the provenance that leads toward such a rectification process, and the procedures involved in collecting personal data from the SoT. Differentiating between proactive and reactive rectifications is helpful for assessments aimed at verifying the controller's commitment to maintaining personal data that is accurate and up-to-date. Since the controller interacts with the SSoT only through requests, those who proactively (e.g., every semester) request the SSoT to rectify personal data exhibit a greater commitment to data quality than those who merely respond to notifications. For the sake of objectivity, in this study, we consider only reactive rectification processing, in which the SSoT reaches the controller to rectify the data.

ities, it is necessary to analyze whether any corrections are required. When the controller collects personal data from the Source of Truth (SoT) for rectification purposes, it temporarily has access to the most accurate version of the data (i.e., information aligned with the SoT). At this point, the controller can initiate procedures to determine whether correction is needed by comparing the newly collected data with the data it already holds. If a discrepancy between these data is identified, a correction is required; otherwise, no correction is needed.

Without delay. The final term requiring clarification concerns the concept of *delay* that controllers must avoid when implementing the necessary changes. What does *without delay* signify in practice? What is the threshold? How should its boundaries be defined? Does it mean a rectification must occur within an hour, or within two minutes? What criteria can serve as a common benchmark to assess whether rectification was performed without delay?

Article 12 (1) lays down the rules for communication between the controller and the data subject to ensure the exercise of the rights established by the GDPR. In particular, this study clarifies the meaning of *without delay* by linking it (via a coordinating conjunction) to a provision specifying that actions must be taken "within one month of receiving the rectification request". The article states:

> The controller shall provide information on action taken on a request under Articles 15 to 22 to the data subject without delay and in any event within one month of receipt of the request.
>
> **GDPR, Article 12, part of point 3**

To our understanding, the term *without delay* is intended to deter lax behavior by controllers in their interactions with data subjects. Nonetheless, this term fails to offer an objective standard for differentiating compliant from non-compliant situations. Thus, for the purposes of this study, we adopt a one-month timeframe for completing actions. We emphasize that this choice is an *operationalization* of the principle based on Article 12 GDPR, not a universal definition of the term: it is the concrete, measurable reading we derive from the only provision in the Regulation that couples "without delay" with an explicit deadline, and it is adopted here solely as the benchmark against which compliance with CQ28 is assessed. A different deployment may legitimately set another threshold, and we expose this value as a replaceable parameter rather than embedding it in the ontology (Section 4.2).

Since the question concerns the duration required to carry out the procedures, our fo-

cus must be solely on the retrospective graph. This duration can be assessed by measuring the time between the commencement of the first procedure and the completion of the last procedure related to this question. The PROV ontology provides the properties **prov:startedAtTime** and **prov:endedAtTime** to represent these temporal boundaries.

4.4.2 GDPRov Extension to address CQ28

(16) have proposed the **bartolini:Rectification** class to represent an action against the controller to rectify their own inaccurate data. (112) represents this as a **ucjcich:Request** started from the data subject towards the controller so that the latter performs the desired action using the informed data. We believe this modeling introduces a counter-intuitive relationship between the data subject and the controller, according to the principle established by the GDPR[8].

Based on the prior analysis, we summarize in Table 4.3 the terms we identify in this question and the derived classes we propose to represent them in the GDPRov ontology.

Concerning the necessary terms discussed in Section 4.4.1, GDPRov already defines several classes to encode them, although it does not include elements to represent the processing for detecting whether a correction is required. Hence, as shown in Table 4.3, we propose the new **:DataComparisonStep** class, extending from **gdprov:DataStep** class and representing the prospective activity of comparing a group of data instances. For retrospective modeling, we introduce **:DataComparisonActivity** class, which is defined as a subclass of **gdprov:DataActivity**. Both of these new classes also inherit from **dpv:Match**, which is used to represent processes that combine, compare, or match data originating from different sources.

To encode the case where a correction is required, so that it becomes possible to query for existing processes related to the data that need rectification and to verify whether the necessary changes have been implemented, we propose the properties **:hasDataDiscrepancy** encoding the outcome of the data discrepancy assessment in a forward and backward-looking fashion, respectively. Such a relation targets a **xsd:boolean** instance, which expresses the result of that comparison. Furthermore, following Pandit's commentary regard-

[8]By representing the data subject requesting the data controller to perform some operation over his/her data, it may reinforce the prior language used before the GDPR, in which the data controller is the real owner of the data and the data subject should ask (permission for) the controller to perform processing over their data. However, as appointed by [[kuner2012european|Kuner]], that is the right copernican reversion the GDPR introduced: by shifting toward the data subject's conception of privacy, we believe that a better representation might be the controller collecting the data from the data subject, as proposed by GDPRov.

Class	Description	Nature
`:DataComparisonStep`	The prospective comparison of the collected data against the stored one. (subclass of `:DataStep`).	Original
`:DataComparisonActivity`	The retrospective comparison of the collected data against the stored one. (subclass of `:DataActivity`).	Original

Table 4.3: Extended classes for GDPRov to address the CQ28. Both classes are *original* to this work: no vocabulary we reviewed provides a concept for the comparison that establishes whether a correction is required. Each specializes two reused parents, a GDPRov class and `dpv:Match`.

Property Domain	Target	Description	Nature
`:hasDataDiscrepancy` `:DataComparisonStep`	`xsd:boolean`	Indicates the outcome of the prospective comparison processing.	Original
`owl:differentFrom` `:PersonalDataEntity`, `:PersonalData`	`:PersonalDataEntity`, `:PersonalData`	The relation establishing the differentiation between collected and stored data.	Reused

Table 4.4: Extended properties for GDPRov to address the CQ28. Only `:hasDataDiscrepancy` is authored here; `owl:differentFrom` is used as defined by OWL, with no local declaration (`xsd:` and `owl:` are prefixes for the ontologies defined in `https://www.w3.org/TR/swbp-xsch-datatypes/` and `https://www.w3.org/TR/owl-features/`, respectively).

ing this issue, we also suggest modeling a detected discrepancy in the provenance graph by employing the property `owl:differentFrom`. All these new properties are detailed in Table 4.4.

4.5 Competency Question CQ29

CQ29: Are retention policies and procedures in place to ensure data are held for no longer than is necessary for the purposes for which they were collected?

When addressing CQ29, Pandit (85) assumes the term “procedures” in the context of this question is a synonymous for data deletion. Also, he assumes that such a procedure should be performed as soon as the data’s original purpose has been fulfilled. His comprehension suggests a causal relationship between fulfilling the data collection purposes and the actual execution of data deletion. By recognizing the challenge of demonstrating this causal

relationship, specifically due to the difficulty of programmatically identifying whether the purpose for which the data was collected has been fulfilled, Pandit scoped this question out of his contribution.

In addition, Pandit highlights the importance of showing that deletion[9] was performed at the correct time. He comments that merely knowing what data was collected, and for what purposes, does not suffice to demonstrate that deletion occurred without delay. Consequently, he suggests we should take retrospective information into account to properly address this question.

In our analysis, Pandit's assumptions connecting the procedures in question to data deletion appear reasonable. Indeed, the phrase "data are held for no longer than is necessary" can be reframed as the conditional statement: "once the data is no longer necessary, the controller must delete or anonymize it." However, we identify an implicit assumption in Pandit's commentary that we do not fully agree with, or at least find overly limiting of the question's aims. From this interpretation, Pandit's comment seems to frame the question solely in terms of retrospective information, disregarding the prospective perspective. This is evident from his use of past forms of the verb and his emphasis on demonstrating that deletion *was performed* at the correct time, rather than focusing on future actions.

In contrast, we argue that this question is more about prospective planning than retrospective verification. The central concern is whether plans are in place to regulate data deletion when the data is no longer needed for its original purpose. We base this interpretation on two key points: first, the sentence in question is phrased in the present tense (data *are* held" rather than data *were* held"), indicating that the focus need not to be limited to past events. Second, if there are any concerns about retrospective events, they have already been addressed by CQ51 (which says "Are personal data systematically destroyed, erased, or anonymised when they are no longer legally required to be retained."). Both CQ29 and CQ51 concern data deletion or anonymization, but they are not redundant: CQ29 concerns *prospective* provenance, i.e., whether a policy or plan for eventual deletion is in place, whereas CQ51 concerns *retrospective* provenance, i.e., whether that plan was actually carried out. This temporal distinction, rather than the underlying obligation, is what separates the two questions.

Therefore, we interpret CQ29 as focusing exclusively on plans for data deletion or anonymization once the purpose for which the data was collected has expired. This interpretation al-

[9] We should also consider cases where deletion is not feasible. In such instances, the GDPR establishes data anonymization as an alternative (see Article 89(1) and Recital 26).

lows reframing the question into the conditional statement discussed in the next section, whose antecedent (the expiration or fulfillment of the purpose) and consequent (the obligation to delete or anonymize the data) define the logical conditions framing the question.

4.5.1 The Terminology in Question CQ29

If... Then. Starting from the question reframing into the conditional statement: "If the purpose for which the data was collected expires or is fulfilled, then the data must be deleted or anonymized"; this structure reflects how EU law operationalises the storage limitation principle (1, arts. 5). Article 17 (1) is explicity in using a conditional language ("[...] where one of the following grounds applies"). Ultimately, this conditional statement is explicitly stated by the European Data Protection Board in (47)[10]. When it comments on the topic of the storage period, it writes: "If personal data is no longer necessary for the purpose of the processing, then it shall by default be deleted or anonymized" (47, item 53).

While GDPRov already provides components to model data-deletion or anonymization procedures, the current model lacks elements to represent the causal relationships between relevant events. This limitation is not confined solely to the GDPRov model; none of the models reviewed in the related literature section (Section 3) provide elements to represent this causal relationship.

Given the existing literature, the Semantic Sensor Network Ontology (SSN)[11] provides a framework for representing relationships among events and processes, including causal connections. By incorporating SSN concepts into GDPRov, the model can explicitly represent the causal relationship between the expiration of data necessity and the initiation of appropriate deletion or anonymization actions.

The Data Purpose Fulfillment. Starting from the antecedent, we address the term concerning data storage conditions, which is closely tied to the *principle of Purpose Limitation* as defined in Article 5(1)(b) of the GDPR. This principle mandates that "personal data shall be collected for specified, explicit, and legitimate purposes and not further processed in a manner that is incompatible with those purposes". Additionally, it is connected to the *Storage Limitation* principle (Article 5(1)(e) of the GDPR), which stipulates that "personal data shall

[10]The European Data Protection Board (EDPB), established by Article 68 GDPR, is an independent body that ensures consistent application of the Regulation across the European Economic Area. It brings together national data protection authorities and the European Data Protection Supervisor, and issues guidelines, recommendations, and best practices to promote uniform enforcement.

[11]Publicly available at `https://www.w3.org/TR/vocab-ssn/`

be kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the personal data are processed."

In Section 4.3, we elicit four scenarios and claim their are enough to represent all the possibilities concerning data storage conditions. For each scenario, we extend the original GDPRov model with a new set of components inherited from well-stablished ontologies: `:TemporalDuration`, `:UntilTimeDuration`, `:UntilEventDuration`, and `:EndlessDuration`. Therefore, with those introduced components, our model is already capable of representing cases where the data purposes have expired, and thus the data is no longer necessary for the original purposes for which it was collected.

The Data Deletion or Anonymization Procedure. In GDPRov, components are provided to address cases where data must be deleted or anonymized. These procedures can be modeled using `:DataErasureProcess` or `:HandleRightOfErasure` components. The steps involved in a prospective plan are specified via `:DataDeletionStep` or `:DataAnonymisationStep`, while retrospective activities are captured by `:DataDeletionActivity` or `:AnonymisationActivity`. The `:AnonymisedDataEntity` component represents data that has been anonymized as the result of some activity. In contrast, prospective data can be represented by `:AnonymisedData`, or `:PersonalData` in conjunction with `:hasAnonymityLevel`. This approach enables the specification of different levels of anonymity, such as `:Anonymised`, `:DeAnonymised`, or `:PseudoAnonymised`, according to the needs of the application.

4.5.2 GDPRov Extension to address CQ29

In light of the analysis presented, Table 4.5 elicits the proposed terms encoded as GDPRov classes and subclasses. This table does not include classes for previously presented terms related to data storage conditions and deletion procedures. Instead, it focuses on the new classes and subclasses introduced by extending GDPRov with SSN concepts to represent the concepts addressed in CQ29.

The first extension consists of aligning the `:Process` class with the `sosa:Procedure` class by making the former a subclass of the latter. This first alignment empowers GDPRov to represent causal relationships between events and processes, as defined by the Semantic Sensor Network Ontology (SSN) concepts.

Second, one of the major approaches to programmatically implementing the causality between two distinct procedures involves leveraging observability mechanisms. An onto-

Class	Description	Nature
`:Process`	By aligning this class with `sosa:Procedure`, we allow it to inherit properties from the Semantic Sensor Network Ontology for representing causality between events and processes (equivalent to `sosa:Procedure`)	Alignment
`:MonitorExpiredData`	Specialize existing process class to represent the specific case of monitoring compliance by checking the fulfillment of data collection purposes (subclass of `:MonitorCompliance`).	Original
`:ExpiredDataObservationStep`	Represent a specific existing step to represent the specific case of monitoring compliance by checking the fulfillment of data collection purposes (subclass of `:DataUsageStep` and equivalent to `sosa:Sensor`).	Original
`:StorageCondition`	The class representing the fulfilment of data collection purposes is made an observable property, so it can serve as the property being observed (equivalent to `sosa:ObservableProperty`).	Alignment

Table 4.5: Classes involved in the GDPRov extension to address the CQ29. The two *alignment* rows introduce no class: `:Process` already belongs to GDPRov and `:StorageCondition` was already reused for CQ8 (Table 4.1); each merely gains an axiom relating it to SSN. Only the two *original* rows add new classes (`sosa:` is a prefix for the ontology defined in `https://www.w3.org/TR/vocab-ssn/`).

logical sight for this approach requires at least three mandatory entities: the observer, the observed, and the act of observing. SSN provides the `sosa:Sensor` class to represent the entity that observes a specific property and the `sosa:ObservableProperty` class to represent the property that is being observed. For the act of observing, SSN provides the `sosa:observes` property.

Since observing data consists of a specific case of data usage, we introduce the `:ExpiredDataObservationStep` class to represent the specific case of `:DataUsageStep` that represents the observer that checks the fulfillment of data collection purposes. This novel class also works as a `sosa:Sensor` in observing the fulfillment of data collection purposes. Additionally, since the fulfillment of data collection purposes is encoded through the `:StorageCondition` class (see Section 4.3), and it plays the role of the observed property, we propose this class should be a subclass of `sosa:ObservableProperty`. Finally, the act of observing is represented by the `sosa:observes` property, which we include in the GDPRov as a new object property.

Finally, since the procedures in question are related to monitoring compliance by check-

Property Domain	Target	Description	Nature
:observes :ExpiredDataObservationStep	:StorageCondition	An object property included in GDPRov for representing the act of observing a particular property (equivalent to sosa:observes).	Reused

Table 4.6: Extended properties for GDPRov to address the CQ29. The property is *reused* from SSN, declared in the GDPRov namespace so that the model can refer to it (`sosa:` is a prefix for the ontology defined in `https://www.w3.org/TR/vocab-ssn/`).

ing whether there are no longer existing data that the original purpose has already been fulfilled, we propose the **:MonitorExpiredData** as a subclass of **:MonitorCompliance** for the specific case of monitoring compliance by checking the fulfillment of data collection purposes.

Table 4.5 and 4.6 outline all these classes and properties necessary to fully represent the concepts addressed in CQ29. Of these, only **:MonitorExpiredData** and **:ExpiredDataObservationStep** are original: the observability machinery itself comes from SSN, and the two alignment axioms simply place GDPRov's own **:Process** and the reused **:StorageCondition** under the SSN roles of procedure and observable property. The originality of the answer to CQ29 therefore rests less on the count of new classes than on the arrangement, since casting purpose fulfilment as an observable property and deletion monitoring as its sensor is what renders representable the causal link that Pandit scoped out and that no model reviewed in Chapter 3 expresses. By incorporating these classes, GDPRov is ready to address emerging aspects of data retention and deletion, ensuring compliance is systematically verified and enforced.

4.6 Competency Question CQ51

CQ51: Are personal data systematically destroyed, erased, or anonymised when they are no longer legally required to be retained?

As anticipated in Section 4.5, this question is the retrospective counterpart to CQ29: whereas CQ29 asks whether a policy or plan for data deletion or anonymization exists, CQ51 asks whether that plan was actually carried out.

Although Pandit has questioned whether this issue can feasibly be addressed using prove-

nance, we argue that provenance information can, in fact, enable a compliance checker to verify whether data have been systematically destroyed, erased, or anonymized "once they are no longer necessary." When such processing operations are captured in the provenance graph, it becomes plausible to affirm the systematic[12] nature of the processing.

First, it is important to clarify the intent of this question. Through queries over the prospective model, it is possible to identify plans that document the procedures for systematically destroying, erasing, or anonymizing personal data. However, the central concern of this question is not only whether such procedures have been identified by the controller, but whether they have been carried out in practice, at the appropriate times. This position becomes clearer when we imagine a hypothetical scenario in which the controller acknowledges the existence of procedures for systematically destroying, erasing, or anonymizing personal data but has not executed them at the appropriate times. This scenario would not be compliant with the question, since it leads to the retention of personal data that is no longer legally required, which violates the condition stated in the question. In other words, the focus is on verifying that the execution of planned operations has been triggered by the expiration of the legal retention period, not merely confirming the presence of a documented plan or policy. Therefore, the competency question requires not only the existence of a plan, but also evidence that it has been followed in practice.

However, taking a direct retrospective approach may be challenging or even unfeasible, in some cases, because the destruction, erasure, or anonymization of data may result in the loss of information needed to trace it. To overcome this potential issue, we might approach the question indirectly by formulating an equivalent inverse query: rather than searching for historical records that confirm expected deletions, we could query for data that remains retained even though it is no longer legally required. In other words, the competency question can be reframed as: "Does any personal data remain after it is no longer necessary?" In such

[12] "What precisely qualifies a procedure as having been performed systematically"? The term *systematic* appears only three times in the GDPR, exclusively in Articles 35 and 37, where it is used as an adjective describing the processing involved in monitoring data subjects or in how the evaluation of personal data should be conducted. The references in Articles 35(1)(c) and 37(1)(b) do not offer further semantic clarification to broaden the understanding of the term. Article 35(1)(a), however, links systematic with *automated* via a nonrestrictive clause, which serves to provide additional information. The mentions in the Recitals similarly add no further insight. Given that *systematic* is not explicitly defined within the GDPR itself, we have consulted external sources for clarification. The EDPB, while rolling as WP29, suggested that systematic "may be interpreted as taking place as part of a general plan; pre-arranged." (12). Guided by this definition, for the purposes of this work, we consider an operation to be systematic if it is performed by automated means (as opposed to manual processes) and follows pre-established steps. This definition aligns well with the definition of prospective provenance, which is a plan for a process to be carried out in the future. There is no need to extend the model to represent the retrospective provenance, since the model already represents the retrospective provenance through the `:Process` class. We prefer to present the analysis of this term within this footnote because its conclusion does not impact the core argument.

a case, it only becomes necessary to demonstrate the absence of any personal data that is no longer legally required to be retained.

This inverse formulation, however, changes the epistemic status of the answer, and the distinction deserves to be made explicit. When the query *returns* a personal-data entity, it exhibits a concrete record of data retained past its lawful period: this is positive evidence of *non*-compliance, and it holds regardless of how complete the rest of the provenance graph is, since a single such record already witnesses the violation. When the query returns *nothing*, by contrast, we may conclude compliance only under a closed-world reading of the graph, that is, only insofar as the provenance record faithfully and completely captures the personal data the controller actually retains. Detecting non-compliance is therefore sound on its own, whereas *proving* compliance rests on the completeness of the provenance record: the absence of a violating record attests compliance only to the extent that everything retained has been recorded. This dependence is not a peculiarity of our approach but a general trait of provenance-based compliance checking; in the GDPR setting it is partly underwritten by Article 30 (1), whose record-of-processing duty is precisely what the controller must satisfy for the provenance graph to be a trustworthy basis for the compliance claim.

By adopting this new perspective, the relevant terminology has largely been addressed in CQ8, which addresses the retention period for personal data. For the remainder of the question, GDPRov already provides suitable constructs to indicate whether data is legally required and whether the prescribed procedures have been followed as planned. Therefore, no further terminological analysis is necessary.

4.7 The Remaining Competency Questions

For the sake of objectivity, this study addresses only some of the uncovered questions in detail. As motivated in Chapter 1, CQ8, CQ28, CQ29, and CQ51 were chosen because together they cover the storage lifecycle of personal data and because they contain the field's hardest open gap — the causal link between purpose fulfilment and deletion that Pandit scoped out and no reviewed model represents. Demonstrating that provenance can address this cluster, the most structurally demanding of the uncovered questions, gives us confidence that provenance will remain relevant for addressing the remaining uncovered questions.

From a broader analysis, we can identify five main categories that can wrap up the remaining uncovered questions. Four of these five categories are questions where provenance can assist with compliance checking. The fifth category encompasses questions that require

a more complex analysis, for which provenance cannot be helpful, in our opinion. Notably, whereas Pandit (85) argued that 13 of the 33 uncovered questions bear no relation to provenance, our broader analysis narrows that set to the six questions of this fifth category, judging the remaining ones amenable to provenance-based compliance checking. It is important to note that these categories are not mutually exclusive, and a question can belong to multiple categories.

The Informative Problems. This first category includes questions CQ21, CQ33, CQ37, CQ38, CQ43, CQ44, and CQ65. These are characterized as "informative problems," which focus on whether an individual has received sufficient information regarding specific aspects of GDPR, such as the rights they possess, the nature of data processing activities, or other relevant information pertaining to their personal data.

Provenance can help address these questions by providing verifiable evidence that such notifications or information events have occurred. For example, for CQ21, which asks whether individuals have been informed of their right to object to a particular processing activity, provenance records may document events such as the presentation of a privacy policy during data subject registration or the transmission of an explicit email informing the individual of this right. These recorded events in the provenance graph serve as direct evidence of notification or information delivery. CQ43, CQ44, and CQ65 also fit within this case.

It is important to note, however, that while provenance can demonstrate the occurrence of such informational events, some questions in this category also require a qualitative or subjective assessment. That is the case of CQ33, which requires additional evaluation about whether the provided information was communicated in a "concise, transparent, intelligible and easily accessible form, using clear and plain language." While provenance can indicate that such notification has occurred, it may not be possible to fully automate the evaluation of these qualitative standards. Instead, provenance can document who made the judgment and when, supporting auditability and accountability. CQ37 and CQ38 also fit within this case.

Simple Demonstrative Problems: The Existence of Activities or Documents. This second category includes questions CQ17, CQ30, CQ39, CQ48, CQ57, and CQ58. These questions are characterized as "simple demonstrative problems," since their focus is on whether specific activities or documents have been carried out or exist. Provenance can help address these questions by providing verifiable evidence that such activities or documents have occurred.

For example, CQ17, which asks whether the organization can respond to SARs (Subject Access Request) within one month, can be addressed through provenance records documenting events such as the initiation of the response process or its completion. These recorded events in the provenance graph serve as direct evidence of the activity or document's existence. CQ48, CQ57, and CQ58 also fit within this case.

CQ30 might be addressed by extending the information from other ontologies that provide the retention rules specified in other laws or regulations. For example, through properties such as `dpv:hasApplicableLaw`, DPV (publicly available at `https://w3id.org/dpv`) enriches the `:Process` with information about the law to which it is subject. CQ39 also fits within this case, requiring additional elements in the model to represent the processing revision.

Decision Demonstrative Problems. This third category encompasses questions CQ40, CQ41, and CQ50, which are particularly complex because they require interpretative assessment based on specific aspects of the law. We refer to these as "decision demonstrative problems." In these instances, provenance can play a supporting role by providing verifiable evidence that such decisions have taken place, specifically, by recording the relevant activities, individuals involved, and supporting documentation.

As an example, CQ40 concerns whether the controller is required to appoint a Data Protection Officer (DPO). Provenance records can document the creation of the decision rationale, the document itself, and the identity of the person authorized to make this determination. Similarly, CQ41 and CQ50 fall within this category, necessitating model extensions to represent both the decision makers and the justifications for their choices.

Conditional Demonstrative Problems. This fourth category includes questions in the *if-then* format. Pandit's comments have included the antecedent sentence in the analysis, which increases their complexity. By assuming that the controller feeds the analysis of the antecedent into the provenance graph, it becomes possible to perform compliance checking supported by the provenance graph.

For example, CQ11 requires the controller to re-seek the individual's consent when the personal data does not meet the GDPR's required standard. Although evaluating whether the personal data meets the required standard under the GDPR is a complex task that cannot be performed solely on the basis of provenance information, provenance can be used to demonstrate that the controller has re-sought the individual's consent when their personal

data does not meet that standard.

Problems Not Covered by Provenance. This fifth and last category comprises those questions that fall outside the boundaries of what provenance can address. Specifically, it includes questions CQ09, CQ32, CQ42, CQ47, CQ49, and CQ52, which encompass a diverse range of topics. These include, for example: identifying actions necessary to ensure compliance, preventing unlawful duplication of data, or escalating certain cases to the DPO (CQ09, CQ32, and CQ42); implementing training and awareness programs (CQ47); assigning individuals to specific roles or tasks (CQ49); and taking particular actions in response to incidents (CQ52).

Chapter 5

Evaluation

The final stage of this extension involves evaluating the proposed ontology. According to the NeOn methodology (108), such an evaluation is defined as "the activity of checking the technical quality of an ontology against a frame of reference." In literature, this *frame of reference* is generally distinguished into two distinct forms of ontology evaluation: *ontology validation* and *ontology verification* (108; 61).

Ontology verification is the process of assessing the technical quality of an ontology by systematically evaluating it against a set of objective criteria to determine whether it has been constructed in accordance with established best practices and methodological standards. In essence, this aspect of evaluation focuses on whether the ontology utilizes appropriate design patterns and adheres to proper modeling conventions. In other words, to verify whether the ontology has been built in the correct manner.

In the literature, a number of tools and methodologies have been developed to facilitate the ontology evaluation process (108; 61; 101). In this study, we employ the OOPS! (OntOlogy Pitfall Scanner!) tool (100)[1]. Based on a catalog of 44 pitfalls and anti-patterns, OOPS! systematically scans ontologies and generates a comprehensive report detailing the identified issues. It is important to clarify that OOPS! covers design and modeling pitfalls only: it evaluates an ontology against a catalog of structural and semantic anti-patterns, rather than performing logical reasoning, so it does not detect logical inconsistencies (e.g., unsatisfiable classes), which would require a description-logic reasoner. With this, we address the verification assessment, ensuring the quality of the proposed ontology modeling.

On the other hand, *ontology validation* concerns determining whether the proposed model accurately represents the real world. This aspect of evaluation focuses on the on-

[1] Available at `https://oops.linkeddata.es`

tology's conceptualization of the domain rather than on its technical and logical considerations, emphasizing the extent to which it captures the intended conceptualization; in other words, whether it represents the right thing. To address this type of assessment, we instantiate the proposed ontology into three real use-case scenarios based on questions CQ8, CQ29, and CQ51.

The remainder of this section is organized as follows. Section 5.1 presents the use case scenario we use to evaluate our proposed extension to GDPRov. It describes the system we use for our evaluation and our privacy policy. The next three sections then explore how GDPRov could help users ensure compliance with specific GDPR policies in this use case. Section 5.2 explores the retention period of personal data (CQ8). Section 5.3 explores retention policies (CQ29). Section 5.4 tackles data deletion (CQ51). Finally, Section 5.5 quantifies the modeling effort and the resulting coverage gain.

5.1 The Use Case Scenario

To validate the proposed model, we use a real-world dataset rather than a purely theoretical one. Using real data enables a thorough evaluation of the ontology's robustness and ensures that any insights gained or limitations uncovered correspond to genuine challenges encountered in actual implementations, thereby minimizing potential biases.

For this purpose, we executed our queries and assessed the proposed model using the SAPOS database. SAPOS[2] is a system designed to support the management of information pertaining to graduate programs. It encompasses multiple functionalities, including managing enrollments, courses, advisement, scholarships, and requirements. The primary aim of SAPOS is to optimize administrative processes within graduate programs, enabling institutions to better manage and monitor students' academic progress.

The main categories of information managed by SAPOS are:

- **Enrollments**: Student registration and course enrollments.
- **Courses**: Course offerings, schedules, and descriptions.
- **Advisement**: Tracking of student-advisor assignments (i.e., which professor advises which students).
- **Scholarships**: Management of scholarship allocations and their statuses.

[2]Available at `https://github.com/gems-uff/sapos`

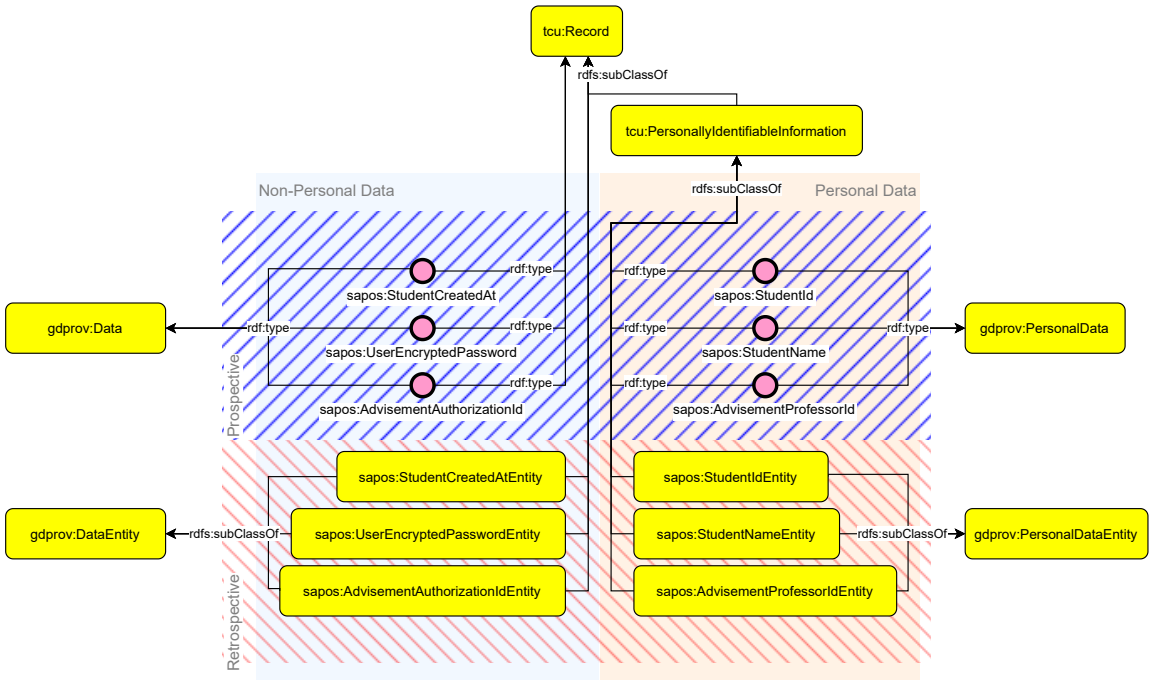

Figure 5.1: Part of the SAPOS alignment with GDPRov and the Trinity College University (TCU) privacy policy ontology. This alignment results in a double representation of each SAPOS data category: a prospective (blue hatched) and a retrospective (red hatched) representation, inheriting it from GDPRov. It also results in a classification effort, which groups the SAPOS data categories into non-personal data (left blue column) or personal data (right red column), aligning with the TCU ontology. Both alignments are given through hierarchical relationships (`rdfs:subClassOf` or `rdf:type`) between TCU concepts (`tcu:Record`, `tcu:PersonallyIdentifiableInformation`), GDPRov classes (`gdprov:Data`, `gdprov:PersonalData`, `gdprov:DataEntity`, `gdprov:PersonalDataEntity`), and specific SAPOS data categories (`sapos:StudentIdData`, `sapos:StudentNameData`, `sapos:StudentCreatedAtData`, etc.). The complete list of SAPOS data categories is omitted for conciseness. This diagram is based on the Graffoo framework specifications.

- **Requirements**: Monitoring of program requirements and student compliance. An example of a requirement could be "a master's student must publish a conference paper before being able to present the master's thesis."

The SAPOS database schema structures this information across 54 tables, which contain a total of 453 attributes[3]. After examining these attributes, we grouped them into three categories, as depicted in Figure 5.1.

SAPOS itself is agnostic with respect to privacy policies. Therefore, to fulfill our objectives, it was necessary to examine a specific graduate program in which SAPOS has been implemented. To the best of our knowledge, some graduate programs at the Universidade

[3]Available at `https://docs.google.com/spreadsheets/d/1pb-fznjb8HLTbzFCD9hCNMNA8yGL5N6Hy7zsPRhLtQ0/edit?usp=sharing`

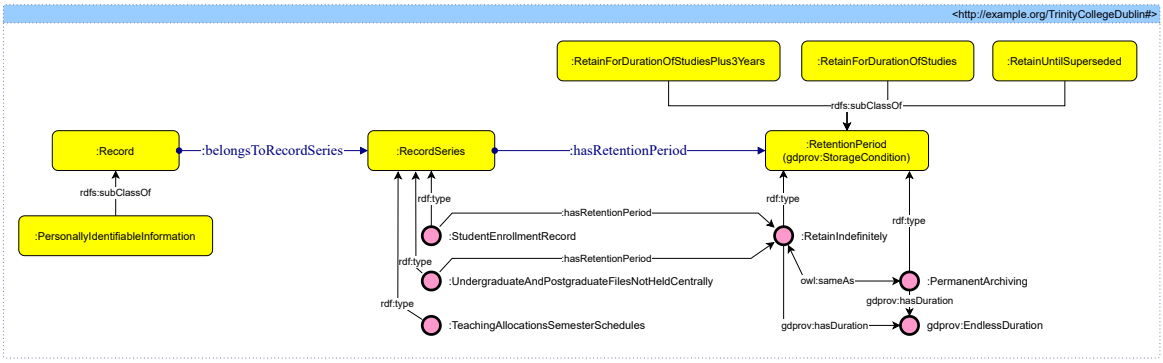

Figure 5.2: Part of the TCU terminology based on Trinity College Dublin's Records Retention Schedule, which uses the terminology of **tcu:Record**, **tcu:PersonallyIdentifiableInformation**, **tcu:RecordSeries** to address the concepts of records, record series, and retention periods, respectively. Within this document, records belong to a record series, which is associated with a retention period. Along with other record series, this document sets the **tcu:StudentEnrollmentRecord**, **tcu:TeachingAllocationsSemesterSchedules**, **tcu:UndergraduateAndPostgraduateFilesNotHeldCentrally**). This diagram was generated using the Graffoo framework.

Federal Fluminense (UFF)[4] utilize SAPOS for course management. However, our search did not identify any ready-to-use, publicly available privacy policies for these programs. Furthermore, UFF is located in Brazil, where personal data processing is regulated by the LGPD [5] (Lei Geral de Proteção de Dados). Although the LGPD exhibits notable similarities to the GDPR (5; 45; 73), we anticipated that differences between the two regulations could potentially influence our results; we examine which differences these are, and how far they bear on generalizing the work to Brazil, in Section 6.

Given these challenges, we opted to consider a privacy policy developed under the scope of the GDPR. Specifically, we examined Trinity College Dublin's (TCD) privacy policy. TCD's privacy framework comprises a family of documents that address different aspects of privacy and data protection. Among these, the Privacy Notice[6] details whose data is collected, the purposes and legal basis for data processing, special categories of data, third-party disclosure, archiving practices, and related aspects. Appendix **??** briefly describes the documents that compose this framework.

It should be acknowledged up front that pairing a Brazilian system with an Irish institution's privacy policy is an artificial arrangement, since no real deployment couples SAPOS with Trinity College Dublin's records-retention schedule. We adopt it as a deliberate, controlled substitution, and it does not weaken what the evaluation sets out to demonstrate.

[4]See more at `https://www.uff.br/`
[5]Available at `https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm`
[6]Publicly available at `https://www.tcd.ie/hr/privacy/`

Both artifacts are real and externally authored: SAPOS is a production system with a fixed schema and genuine data categories, and the TCD documents form a publicly published, GDPR-scoped privacy framework with actual record series and retention periods. Because neither was written to fit our model, the alignment exercises the proposed constructs against evidence we did not control, which a policy fabricated to match the ontology would not have done. Moreover, the evaluation is an ontology validation in the representational sense (Section 5.5) rather than an empirical audit of how a given organization behaves; what it must establish is that the model can represent and query the retention, monitoring, and deletion information a GDPR policy prescribes, and that capability does not depend on the system and the policy sharing a jurisdiction. What the pairing does not warrant is any conclusion about SAPOS's own compliance or about Brazilian deployments under the LGPD, a boundary on external validity we return to in Section 6.

Building upon this family of documents, we constructed a domain-specific ontology aligned with the GDPRov model, namely the Trinity College University (TCU) ontology, so named after the institution whose privacy framework it encodes. Figure 5.2 illustrates the key elements of the TCU ontology, which we summarize as follows:

- **tcu:Record**: Within the TCU vocabulary, all personal data is denoted as a "record". This concept can have the status of "permanently valuable record" (**tcu:PermanentlyValuableRecord**), "inactive record" (i.e., **tcu:InactiveRecord**), "semi-active record" (**tcu:SemiActiveRecord**), or "active record" (**tcu:Active\-Record**). Also record can assume the form of "personally identifiable information (**tcu:ActiveRecord**).

- **tcu:RecordSeries**: According to the Records Management Policy[7], each record is aggregated within a "record series". This relationship is captured via the **tcu:belongsToRecordSeries** property, which may refer to entities such as **tcu:TeachingAllocationsSemesterSchedules**, **tcu:StudentEnrollmentRecord**, and **tcu:UndergraduateAndPostgraduateFilesNotHeldCentrally**. Since these entities cannot assume more specific form, we represent them as instances rather then as classes.

- **:RetentionPeriod**: The Record Retention Schedule[8] sets the retention period for each record series. In our ontology, this class is modeled as a subclass of **gdprov:StorageCondition**. Instances can represent, for example, retention for the duration of studies

[7] Publicly available at https://www.tcd.ie/media/tcd/about/policies/pdfs/160713-Records-Management-Policy-website.pdf

[8] Publicly available at https://www.tcd.ie/media/tcd/about/policies/pdfs/160713-Records-Management-Policy-Records-Retention-Schedule-website.pdf

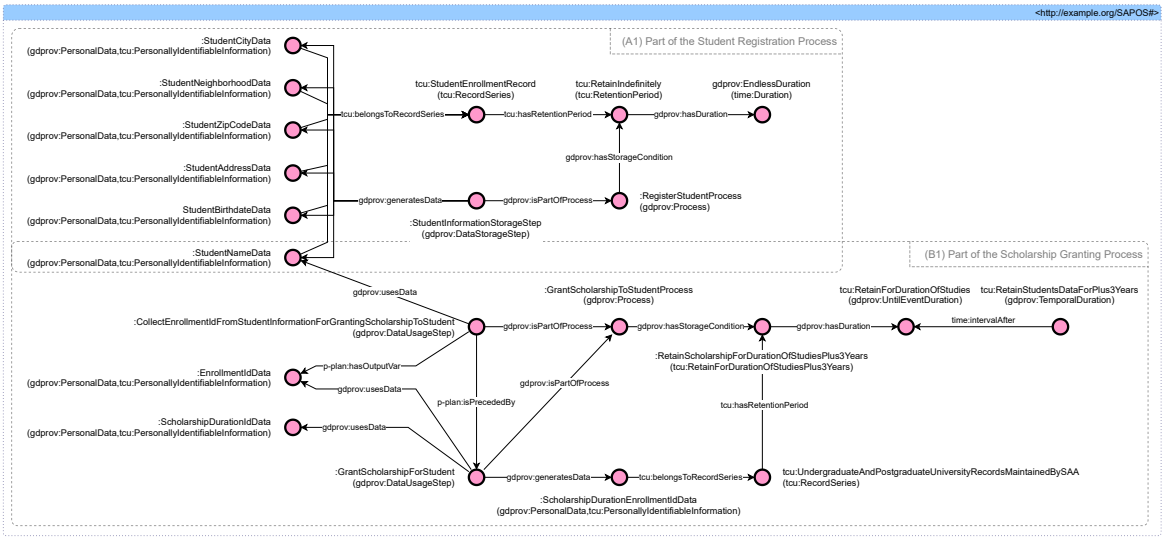

Figure 5.3: Graphical representation of the use case scenario 1: student registration (A1) and scholarship granting process (B1).

plus three years (`tcu:RetainForDurationOfStudiesPlus3Years`), retention for the duration of studies only (`:RetainForDurationOfStudies`), or indefinite retention (`:Retain\-ForDurationOfStudies`), which specifies an unlimited duration.

5.2 Use Case Scenario 1: Student Registration and Scholarship Granting

This first use case scenario describes the plans (i.e., the prospective graphs) for the student registration and scholarship granting process according to the SAPOS implementations under the TCU privacy policy. Figure 5.3 graphically instantiates the elements (e.g., steps, data categories, etc.) involved in this scenario using the GDPRov ontology. It wraps these elements within two dashed boxes named in accordance to the represented processes: "(A1) Student Registration Process" and "(B1) Scholarship Granting Process". Both boxes solely offer visual guidance, while the actual semantics comes from the GDPRov classes and properties reused throughout this evaluation.

In SAPOS, the student registration process, which `:RegisterStudentProcess` represents, is composed of the step of data collection, namely `:DataCollectionStepForStudentRegistration`, during which the university collects personal details such as the student's name, date of birth, address, and city. According to the TCU privacy policy, these details belong to *the student enrollment record series*, represented by `tcu:StudentEnrollmentRecord`, which is kept for an unlimited period of time, as encoded by `tcu:RetainIndefinitely`.

```
PREFIX : <https://w3id.org/GDPRov#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

SELECT DISTINCT ?dataCategory ?storageCondition ?duration
WHERE {
    # ?dataCategory rdf:type tcu:PersonallyIdentifiableInformation .
    ?dataCategory rdf:type/rdfs:subClassOf* :PersonalData .

    ?step :generatesData ?dataCategory .
    ?step :isPartOfProcess/:hasStorageCondition ?storageCondition .

    ?storageCondition :hasDuration ?user_duration .
    ?user_duration rdf:type?|rdfs:subClassOf* ?duration .
    ?duration rdf:type|rdfs:subClassOf :Duration .
}
ORDER BY ?dataCategory ?storageCondition ?duration
```

Listing 5: The encoded version of the question "For each category of personal data, list the period for which the data will be retained e.g. one month? one year?" in SPARQL.

This process is depicted in section (A1) of Figure 5.3.

Later, when this same student obtains a scholarship, SAPOS registers it in the system following two sequential steps: first, it searches for the student's enrollment ID using the student's name (`:CollectEnrollmentIdFrom\-StudentInformationForGrantingScholarshipToStude`) and then it assigns an existing scholarship[9] to this student (`:GrantScholarshipForStude⌋nt`) using his enrollment ID. At the end, SAPOS represents this new scholarship assignment by creating a new scholarship duration record associated with this enrollment ID (`:Schola⌋rshipDurationEnrollmentIdData`) in the system.

According to the TCU privacy policy, this scholarship assignment information belongs to *the undergraduate and postgraduate university records maintained by the Student Administration and Affairs (SAA)* record series[10], which `tcu:UndergraduateAndPostgraduateU⌋niversityRecordsMaintainedBySAA` represents and which is kept for the duration of the student's studies plus three additional years, as encoded by `tcu:RetainForDurationOfSt⌋udiesPlus3Years`. All these elements are depicted in section (B1) of Figure 5.3.

From this first scenario, we can systematically retrieve the categories of personal data

[9] In SAPOS, the same scholarship can be assigned to multiple students (one at a time) along its existance. Each student holds the scholarship for a given non-overlaping period.

[10] Publicly available at `https://tinyurl.com/24h5dmaj`, page 7.

```
1 'http://example.org/SAPOS#StoredScholarshipDurationEnrollmentIdData',
  ↪ 'http://example.org/SAPOS#RetainScholarshipForDurationOfStudiesPlus3
  ↪ Years',
  ↪ 'https://w3id.org/GDPRov#UntilEventDuration'
2 'http://example.org/SAPOS#StoredStudentAddressData',
  ↪ 'http://example.org/TrinityCollegeDublin#RetainIndefinitely',
  ↪ 'https://w3id.org/GDPRov#EndlessDuration'
3 'http://example.org/SAPOS#StoredStudentBirthdateData',
  ↪ 'http://example.org/TrinityCollegeDublin#RetainIndefinitely',
  ↪ 'https://w3id.org/GDPRov#EndlessDuration'
4 'http://example.org/SAPOS#StoredStudentCityData',
  ↪ 'http://example.org/TrinityCollegeDublin#RetainIndefinitely',
  ↪ 'https://w3id.org/GDPRov#EndlessDuration'
5 'http://example.org/SAPOS#StoredStudentNameData',
  ↪ 'http://example.org/TrinityCollegeDublin#RetainIndefinitely',
  ↪ 'https://w3id.org/GDPRov#EndlessDuration'
6 'http://example.org/SAPOS#StoredStudentNeighborhoodData',
  ↪ 'http://example.org/TrinityCollegeDublin#RetainIndefinitely',
  ↪ 'https://w3id.org/GDPRov#EndlessDuration'
7 'http://example.org/SAPOS#StoredStudentZipCodeData',
  ↪ 'http://example.org/TrinityCollegeDublin#RetainIndefinitely',
  ↪ 'https://w3id.org/GDPRov#EndlessDuration'
```

Listing 6: The result of the SPARQL query shown in Listing 5 for the use case scenario 1.

and the retention period for each category by executing the SPARQL query provided in Listing 5. This information matches with the expected answer to CQ8.

Because this question requires listing the categories of personal data, it is necessary to connect the SAPOS ontology to the proposed model (i.e., GDPRov) by specifying which attributes constitute Personal Data. This is accomplished by ensuring that **sapos:PersonalData** is a subclass of **gdprov:PersonalData**, which allows us to accurately identify the relevant personal data categories. Once each personal data attribute is linked to its corresponding record series, as illustrated in Figure 5.1, and each record series is associated with a storage condition (**tcu:RetentionPeriod**), we can determine the retention policy for each data category.

As shown in Listing 5, the query retrieves the data categories and their corresponding storage condition, and detailed information about the associated retention duration (see line 6). A data category is identified as any instance of any subclass of **gdprov:Data** (line 7). Once the data category is identified, the query traverses through the related **gdprov:Step** and **gdprov:Process** to determine the appropriate storage conditions (lines 9-10).

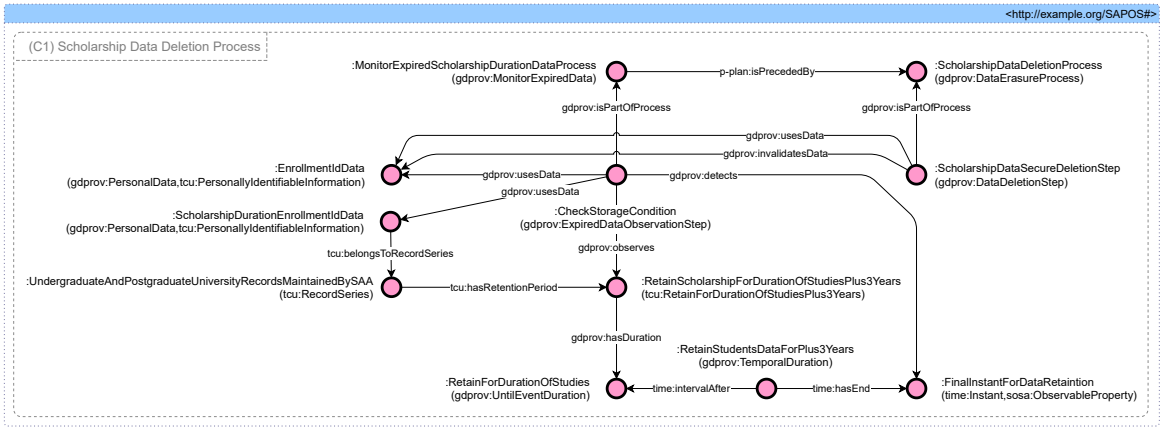

Figure 5.4: Graphical representation of the use case scenario 2: monitoring expired scholarship duration data and triggering the archival/anonymisation process.

Finally, starting from the storage condition, the query extracts all relevant details regarding the duration (lines 12-14).

The query result is depicted in Listing 6, which presents the expected result for CQ8 for this scenario: the data categories identified are *enrollment ID* and *scholarship duration enrollment ID* (both of which will be kept for the duration of the student's studies plus three additional years), and *student address*, *student birthdate*, *student city*, *student name*, *student neighborhood*, and *student zip code* (which will be kept for an unlimited period of time).

Finally, it should be emphasized that the query remains independent of any particular use-case model, as it does not need to reference any element from the specific use-case vocabulary. As a result, the queries developed are broadly generalizable and can be applied to multiple scenarios.

5.3 Use Case Scenario 2: Scholarship Retention Monitoring and Archival

This second use case scenario details how SAPOS, operating under the TCU privacy policy, enforces the retention rules associated with the scholarship granting plan. Figure 5.4 extends the use case introduced in Section 5.2 by following the personal data produced during the scholarship assignment until the moment the purpose for which such data was collected is no longer necessary.

Currently, SAPOS lacks automated mechanisms to delete data that are no longer necessary for their original intended purposes. Instead, this task must be performed manually

by the SAPOS administrator, who verifies the storage condition only when a deletion is explicitly requested. To improve compliance, we propose that SAPOS should periodically (for example, at the end of each semester) send the administrator a list of all students who graduated more than three years ago (according to the TCU privacy policy). The administrator would then be responsible for manually checking whether the storage condition has expired for each case and executing the deletion process if appropriate.

Regardless of the approach taken, SAPOS always use the scholarship duration ID to check whether the storage condition has indeed expired (Figure 5.4 focus specifically on this verification process). This is accomplished by searching for the student's enrollment information (id, situation, graduation date, etc.) using the enrollment ID. Once found, the administrator checks (`:observes`) whether the storage condition associated with the scholarship duration ID has indeed expired (`:CheckedIfAlicesStorageConditionHasFinished`).

According to the TCU privacy policy, scholarship information should be kept for the duration of the studies plus three additional years and then securely deleted (`:RetainScholarshipForDurationOfStudiesPlus3Years`). Once the administrator detected (`:detects`) that the storage condition has indeed expired, (s)he starts the deletion process that erases the scholarship information from the SAPOS database (`:ScholarshipDataDeletionProcess`), which invalidates the original scholarship information. This process is depicted in Figure 5.4, section (C1).

By applying the SPARQL query shown in Listing 7 to this second use case scenario, we can automatically assess how long the controller complies with CQ29 by checking the existence of the monitoring and enforcement procedures for all personal data categories.

The SPARQL query depicted in Listing 7 selects distinctly the personal data categories and their corresponding retention policy, monitoring process, and enforcement process (line 6). We highlight that the monitoring and enforcement processes must be optional (lines 15-21 and 23-27, respectively) so that we can capture the cases where the data is not monitored or enforced.

The result of this query is depicted in Listing 8, which presents the expected result for CQ29 for this second use case scenario. Note that for this depicted scenario, the personal data category *enrollment ID* is not monitored or enforced. This is because we do not have specified these procedures for this personal data category in the use case scenario. On the other hand, the personal data category *scholarship duration enrollment ID* is monitored and enforced. These results illustrate well the case where we can verify the compliance with CQ29.

```
PREFIX gdprov: <https://w3id.org/GDPRov#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

SELECT DISTINCT
    ?personalData ?retentionPolicy ?monitoringProcess
    ↪ ?enforcementProcess
WHERE {
    ?personalData rdf:type/rdfs:subClassOf* gdprov:PersonalData .

    ?collectionStep gdprov:isPartOfProcess ?collectionProcess ;
                    gdprov:generatesData ?personalData .
    ?collectionProcess rdf:type/rdfs:subClassOf* gdprov:Process ;
                    gdprov:hasStorageCondition ?retentionPolicy .

    OPTIONAL {
        ?monitoringProcess rdf:type/rdfs:subClassOf*
        ↪ gdprov:MonitorExpiredData .
        ?observationStep rdf:type/rdfs:subClassOf*
        ↪ gdprov:ExpiredDataObservationStep ;
                        gdprov:isPartOfProcess ?monitoringProcess ;
                        gdprov:observes ?retentionPolicy ;
                        gdprov:usesData ?personalData .
    }

    OPTIONAL {
        ?enforcementStep rdf:type/rdfs:subClassOf* ?stepType ;
                        gdprov:isPartOfProcess ?enforcementProcess ;
                        gdprov:invalidatesData ?personalData .
    }
    VALUES ?stepType { gdprov:DataDeletionStep
    ↪ gdprov:DataAnonymisationStep }
}
ORDER BY ?personalData ?retentionPolicy
```

Listing 7: The encoded version of the question CQ29: “Are retention policies and procedures in place to ensure data are held for no longer than is necessary for the purposes for which they were collected?” in SPARQL.

Finally, it should also be emphasized that the query remains independent of any particular use-case model, as it does not need to reference any element from the specific use-case vocabulary. As a result, the developed queries are broadly generalizable and can be applied to multiple scenarios.

```
 'http://example.org/SAPOS#EnrollmentIdData', 'http://example.org/SAPOS
  #RetainScholarshipForDurationOfStudiesPlus3Years', 'None',
  'None'
 'http://example.org/SAPOS#ScholarshipDurationEnrollmentIdData',
  'http://example.org/SAPOS#RetainScholarshipForDurationOfStudiesPlus3
  Years',
  'http://example.org/SAPOS#MonitorExpiredScholarshipDurationDataProce
  ss',
  'http://example.org/SAPOS#ArchiveScholarshipDataProcess'
```

Listing 8: The result of the SPARQL query shown in Listing 7 for the use case scenario 2.

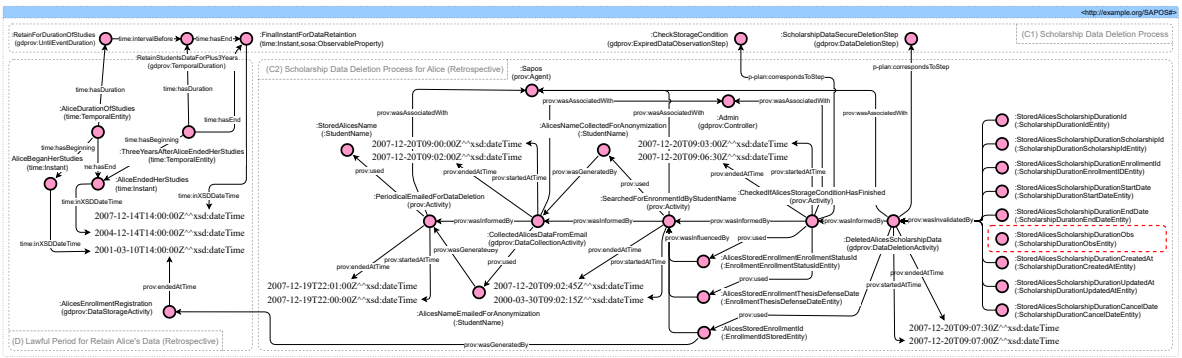

Figure 5.5: Graphical representation of scholarship data lifecycle. Three dashed boxes represent: At the top, **(C1) Scholarship Data Deletion Process** depicts part of the planned deletion process;**(D) Lawful Period for Retain Alice's Data (Retrospective)** - shows the timeline during which Alice's scholarship data must be legally retained, spanning from her study completion through the required retention period; and **(C2) Scholarship Data Deletion Process for Alice (Retrospective)** - illustrates the actual execution of data deletion procedures for Alice's scholarship information after the retention period expired. The red dashed box highlights that the observation data was not deleted on this example.

5.4 Use Case Scenario 3: Data Deletion

Use Case Scenario 5.3 has described the planned process for scholarship data deletion processes (C1). In contrast, this third scenario (illustrated in Figure 5.5) details the real-world case of Alice, who received a scholarship in 2000 and, in 2007, must be deleted according to the TCU privacy policy (C2).

SAPOS currently does not automatically delete data that are no longer necessary for their original purposes; rather, deletion is performed manually by the SAPOS administrator, who verifies the storage condition only when a deletion is explicitly requested. To overcome this limitation, we suggest that SAPOS should periodically (e.g., after each semester) send the administrator an email listing all students who graduated three years prior, according to TCU

privacy policy. This would enable the administrator to review the list and manually execute the deletion process for any data whose storage condition has expired. This use case presumes and describes this recommendation.

On December 19, 2007, SAPOS sent an email to the SAPOS administrator requesting the deletion of, inter alias, Alice's scholarship information (**:EmailedRequestingHerDataDeletion**). At 9:00 a.m. of the following day, the administrator acted upon this request by retrieving Alice's name from the email and using it to locate her enrollment information (such as enrollment ID, status, and thesis defense date). Upon identifying the relevant records, the administrator determined that the graduate program no longer had a legal basis for retaining Alice's scholarship duration data (**:CheckedIfAlicesStorageConditionHasFinished**). Consequently, the administrator executed the data deletion process (**:ScholarshipDataDeletionProcess**), thereby invalidating Alice's scholarship duration information. This sequence is depicted in Figure 5.5, section (C2).

Figure 5.5 (D) illustrates the retrospective timeline indicating the period during which Alice's scholarship duration data was legally required to be retained until its expiration. It begins by showing the date Alice completed her studies (i.e., 2004-12-14), followed by the subsequent three-year period (i.e., until 2007-12-14) during which her scholarship data needed to be retained in accordance with the TCU privacy policy.

In this scenario, CQ51 examines whether all personal data are systematically destroyed, erased, or anonymized once retention is no longer legally required. As discussed in Section 4.6, CQ51 should be evaluated indirectly: due to potential loss of records needed to confirm deletions, the focus shifts to identifying any data that remains stored beyond its lawful retention period. Thus, rather than searching for historical evidence of completed deletions, the key question becomes: "Does any personal data remain that is no longer legally necessary to retain?" In this approach, demonstrating compliance requires showing that no such data exists.

For the sake of demonstration, as depicted in Figure 5.5 (C2), the SAPOS administrator have not deleted the observation data (**:StoredAlicesScholarshipDurationObs**) related to Alice's scholarship duration, although all other related data was properly erased. Consequently, this scenario does not meet the requirements of CQ51, as some personal data remains that should have been deleted after its retention period expired.

Following the interpretation of CQ51, we encode this question by examining the retrospective dataset for any personal data entities whose retention period has expired without

documented invalidation (i.e., neither deletion nor anonymization has occurred).[11] Listing 9 illustrates this inverse formulation of CQ51. The query begins by selecting storage conditions linked to any personal data entity ever persisted (`:DataStorageStep`) (lines 3–7). These storage conditions indicate the policy-defined duration category per GDPRov (lines 9–10). For the use case described in this Section, the duration types are `gdprov:UntilEventDuration` and `gdprov:TemporalDuration`. Lines 12–13 establish the link between the stored data, the storage condition, and the date on which retention is no longer required (`:FinalInstantForDataRetaintion`). Finally, lines 15–17 apply a filter to exclude data already invalidated (either deleted or anonymized), ensuring that the query returns only personal data that should have been deleted, but remains present.

Listing 10 presents the outcome of executing this query on the retrospective trace illustrated in Figure 5.5. The returned tuple indicates that `:StoredAlicesScholarshipDurationObs` remains stored, despite its associated `gdprov:TemporalDuration` having ended on December 14, 2007. Since Lines 15–17 of Listing 9 filter out all personal data entities already invalidated (for example, by `gdprov:DataErasureProcess`), the continued presence of `:StoredAlicesScholarshipDurationObs` constitutes direct evidence that the scenario depicted in Figure 5.5 fails to comply with CQ51.

This outcome illustrates the asymmetry inherent in the inverse formulation, worth restating with the concrete result in hand. The query *detected* non-compliance: the returned entity is a witness of data retained beyond its lawful period, and this conclusion stands independently of whether the rest of the graph is complete. Had the query instead returned no rows, we could have read that as compliance only under the assumption that the provenance graph records every personal-data entity the controller retains, since an incomplete record would produce an equally empty result. The evaluation therefore soundly exposes violations, whereas a clean result certifies compliance only as far as the provenance record is complete, as discussed in Section 4.6.

5.5 Quantitative Analysis

Although the preceding scenarios validate the model demonstratively, we also quantify the modeling effort and the coverage gain. Table 5.1 reports, per competency question, how many constructs the extension contributes and of what kind, following the original, reused,

[11] The specific encoding of this query may differ depending on the five types of storage condition durations presented in Section 4.3, as well as their possible combinations. However, for the scope of this study, we focus on the use of `gdprov:UntilEventDuration` and `gdprov:TemporalDuration`, as reflected in the TCU privacy policy. This combination addresses the most frequent cases observed in privacy policies, according to (80).

```
SELECT ?storedDataEntity (GROUP_CONCAT(?durationType; SEPARATOR=", ") AS
 ↪  ?durationTypes) ?expectedEndDateValue
WHERE {
    ?storedDataEntity rdf:type/rdfs:subClassOf+ :PersonalDataEntity .
    ?storedDataEntity prov:wasGeneratedBy/pPlan:correspondsToStep ?step .

    ?step rdf:type/rdfs:subClassOf* :DataStorageStep .
    ?step :isPartOfProcess/:hasStorageCondition ?storageCondition .

    ?storageCondition :hasDuration/rdf:type/rdfs:subClassOf*
     ↪  ?durationType .
    ?durationType rdfs:subClassOf :Duration .

    ?storageCondition :hasDuration/time:hasEnd/time:inXSDDateTime
     ↪  ?expectedEndDateValue .
    ?storageCondition
     ↪  :hasDuration/^time:hasDuration/time:hasBeginning/time:inXSDDateT⌋
     ↪  ime/^prov:endedAtTime/^prov:wasGeneratedBy/^prov:used
     ↪  ?deletionActivity .

    FILTER NOT EXISTS {
        ?storedDataEntity prov:wasInvalidatedBy ?deletionActivity .
    }
}
GROUP BY ?storedDataEntity ?expectedEndDateValue
ORDER BY ?storedDataEntity ?expectedEndDateValue
```

Listing 9: The encoded version of the question CQ51: “Are personal data systematically destroyed, erased, or anonymised when they are no longer legally required to be retained?” in SPARQL.

```
'http://example.org/SAPOS#StoredAlicesScholarshipDurationObs',
 ↪  'https://w3id.org/GDPRov#TemporalDuration,
 ↪  https://w3id.org/GDPRov#UntilEventDuration', '2007-12-14T14:00:00'
```

Listing 10: Result of Listing 9 for Alice’s deletion scenario.

and alignment distinction introduced in Section 4, together with the external vocabularies drawn upon.

The breakdown makes the balance of the contribution explicit. The extension declares 11 classes and 4 properties in the GDPRov namespace and adds 2 alignment axioms, but only 4 classes (**:DataComparisonStep**, **:DataComparisonActivity**, **:MonitorExpiredDa⌋ta**, and **:ExpiredDataObservationStep**) and 1 property (**:hasDataDiscrepancy**) are orig-

CQ	Original cl. / pr.	Reused cl. / pr.	Align.	Reused vocabularies
CQ8	0 / 0	7 / 2	0	DPV, Time
CQ28	2 / 1	0 / 1	0	DPV (`dpv:Match`), OWL, XSD
CQ29	2 / 0	0 / 1	2	SSN/SOSA
CQ51	0 / 0	0 / 0	0	reuses the CQ8 constructs (inverse-query formulation)
Total	**4 / 1**	**7 / 4**	**2**	—

Table 5.1: Modeling effort and coverage of the proposed extension, broken down by the nature of each construct (cl. = classes, pr. = properties) and computed from the class and property tables of Section 4. *Original* constructs have no counterpart in the reviewed vocabularies; *reused* ones carry the semantics of an external vocabulary; *alignments* add an axiom to a class that already exists, introducing no new class. Constructs are counted once, at the question that introduces them, so `:StorageCondition` is counted under CQ8 and appears under CQ29 only as an alignment. Coverage rises from 32 (85) to 36 of the 65 GDPR Guidance questions (+4; +12.5% over the prior model).

inal, in the sense of having no counterpart in the vocabularies we reviewed; everything else reuses DPV, Time, SSN/SOSA, OWL, and XSD terms. This proportion is a design choice rather than a shortfall: NeOn prescribes reusing an existing resource whenever one fits (107), so the authored part concentrates in the few concepts the community vocabularies do not supply, while the rest of the value lies in the alignment axioms that let those vocabularies operate inside a provenance model, which none of them does on its own. CQ51 required no element of either kind, being answered by reusing the CQ8 constructs under an inverse-query formulation. At the corpus level, the extension raises the number of GDPR Guidance questions answerable by GDPRov from 32 (85) to 36, a 12.5% increase over the prior model.

We stress that both figures reported in Table 5.1 are derived from the extension itself (its class and property tables) rather than from new experiments, so they quantify modeling effort and coverage gain rather than experimental results; this caveat is stated here, before the table, to prevent any misinterpretation of the numbers below.

It is worth positioning this evaluation explicitly. The validation reported here is, by design, an ontology *validation* in the NeOn sense (108): it establishes that the model represents the intended domain by answering the target questions over a real instantiation, rather than a performance or comparative benchmark. We deliberately keep the modeling-effort figures descriptive: they make the cost discussed in Section 4.2 concrete (classes and properties added, vocabularies reused) without claiming statistical generalization. Query performance and scalability are out of the scope of this validation and are treated as future work in Section 6, consistent with the broader field, where comparable models are likewise validated for

feature existence rather than performance (85; 65; 23).

Beyond this demonstrative validation, we submitted the final extended model to OOPS! (OntOlogy Pitfall Scanner!) (100), a methodology-independent tool that scans an ontology for common modeling pitfalls and classifies them by importance. As noted in the opening of this chapter, the scan covers design and modeling pitfalls only, not logical inconsistencies, which would require a reasoner. The extended GDPRov passed the scan with no critical or important pitfalls detected, indicating that the introduced classes and properties are free of the major structural modeling errors catalogued by the tool, such as polysemous elements or synonyms modeled as distinct concepts.

Chapter 6

Conclusion

This study has demonstrated the feasibility of extending the GDPRov ontology to address previously uncovered GDPR compliance questions, expanding the corpus of official questions that can be answered by automated techniques via provenance model. Our results provide both theoretical and practical insights into the automation of GDPR compliance checking through provenance and semantic web technologies.

Our proposed extension integrates concepts from the Data Privacy Vocabulary (DPV) and Time ontologies to represent retention periods and duration types (CQ8). We introduce classes for data comparison procedures and discrepancy detection (CQ28). We incorporate Semantic Sensor Network Ontology (SSN) concepts to model causal relationships between data purpose fulfillment and deletion procedures (CQ29). We demonstrate retrospective compliance verification by querying for data stored beyond the retention periods (CQ51). Additionally, our analysis of the remaining 29 uncovered questions revealed five distinct categories. These categories suggest systematic means for future extensions.

By performing a systematic literature review, we have mapped five distinct methodological categories that the literature has employed to address GDPR compliance issues. These include *CQ-driven approaches based on official check-lists*, which offer superior traceability to regulatory expectations; *CQ-driven approaches using author-defined questions*, providing flexibility for specific compliance scenarios; *directive-driven approaches grounded directly in GDPR articles*; *community-driven techniques leveraging established vocabularies and ontologies*; and *interpretation-based approaches that derive requirements from GDPR text analysis*. Each category presents distinct advantages in terms of traceability, flexibility, community adoption, and level of complexity for the stakeholders to address specific compliance issues.

Since this study adopted a CQ-driven approach, using official check-lists to evolve the

model, we can arguably state that this contribution advances the state of the art. CQ-driven approaches present the advantages of making the traceability easier, and covering the regulatory expectations and literature limitations. Compared to the literature, our approach enables coverage of 4 additional questions (CQ8, CQ28, CQ29, CQ51), representing a 12.5% increase in coverage. Additionally, by extending the new components from well-established community-driven vocabularies (DPV, Time ontology, SSN), our contribution takes advantage of high consensus and wide community adoption. The extended ontology and the complete set of use-case-independent SPARQL query patterns are publicly available at `https://github.com/danielpcampagna/GDPRov`.

Taken together, these results allow us to revisit the research question and objective posed in Chapter 1. The **RQ** asked to what extent data provenance models can answer the official GDPR compliance questions that current models leave uncovered. The four newly answered competency questions, validated over a real database and a real privacy policy through use-case-independent SPARQL queries, answer it in measured terms: an existing provenance model can be extended to cover four of the officially posed obligations the state of the art had left uncovered, though not the entire remainder. Correspondingly, the **RO** of proposing an extension to an existing data provenance model able to represent, query, and validate the information required to answer CQ8, CQ28, CQ29, and CQ51 was achieved, raising the coverage of the GDPR Guidance for SMEs from 32 to 36 of its 65 questions. These results support the representational feasibility and adequacy of the extended model for the four addressed questions; they do not, by themselves, establish that the model generalizes across other organizations, jurisdictions, or scales, a boundary the limitations below discuss in detail.

Despite these advances, there are still some limitations that need further discussion and evaluation. These limitations open the door for future contributions and research:

Limited Scope and Evaluation. Our validation employs a single real-world dataset (SAPOS) with one privacy policy (Trinity College Dublin) across three scenarios. That single pairing is moreover cross-jurisdictional by construction, coupling a Brazilian system with an Irish policy; we justify this in Section 5.1 as a controlled substitution that leaves the representational demonstration intact, but a future multi-policy validation should nonetheless replace it with natively matched system-policy pairs. Although this validation is limited to a small number of use cases (aligning with current practices in the field, where prominent works such as (85; 105; 28; 112) similarly validate with single or limited use cases) it represents a field-wide limitation. To strengthen confidence in the model's generalizability, future work should stress-test it against intentionally diverse datasets and privacy policies from different

jurisdictions. The cross-jurisdiction compliance instruments surfaced in Section 3.2 (Table 1.1), among them the AEPD, ICO, CNIL, NIST, BayLDA, and ANPD checklists, provide ready external references for such a multi-policy, multi-jurisdiction validation, allowing the model to be tested against questions phrased independently of the Irish DPC checklist and against a non-GDPR regime (the LGPD). The absence of negative test cases and adversarial scenarios in current GDPR ontology evaluation methodologies remains an open challenge for the entire research community.

Additionally, 29 of the 65 questions remain unaddressed. Rather than tackling them ad hoc, we propose a concrete roadmap: extend coverage category by category following the five-category map of Section 4.7, beginning with the four categories we identify as provenance-amenable (which together account for 23 of the 29 questions) and treating the fifth, non-provenance category (the remaining 6 questions) as an explicit boundary of the approach. Proceeding by category rather than question by question lets the constructs introduced for one question be reused across its category, turning the remaining coverage into a structured, prioritizable agenda.

Furthermore, beyond the automated pitfall scan with OOPS! reported in Section 5.5, our evaluation focuses primarily on competency question answering rather than the comprehensive multi-dimensional framework recommended by the NeOn methodology (107), which includes structural metrics (e.g., relationship richness, inheritance depth) and task-based evaluation with domain experts. Future work should incorporate these complementary evaluation dimensions to provide a more holistic assessment of the ontology's quality.

Generalization Beyond the GDPR: The LGPD Case. Because the SAPOS system is Brazilian, a natural question is how far the results transfer to Brazil's General Data Protection Law (LGPD) (24), whose differences from the GDPR we anticipated (Section 5.1) could influence the outcome. The transfer is plausible at the structural level but is not automatic, and it helps to separate what carries over from what does not. What carries over is the representational core, since the essential constructs of the extension (Section 4.2) are derived from data-protection *principles* rather than from the wording of any single statute. The LGPD shares the principles our four questions operationalize: storage limitation and the termination of processing once its purpose is met (arts. 15 and 16), the data-quality requirement that personal data be accurate and up to date (art. 6, item V), and the data-subject rights of correction and elimination (art. 18, items III and VI). Because the duration, comparison, and observability constructs encode these principles and not GDPR article numbers, a controller under the LGPD could instantiate the same model without redesigning its class structure.

What does not carry over automatically are the *interpretive parameters* that the model deliberately isolates as replaceable (Section 4.2), two of which are GDPR-specific. First, our operationalization of "without delay" as one month follows GDPR Article 12(3), which fixes a single deadline across data-subject requests; the LGPD sets its response windows differently (for confirmation and access, art. 19 prescribes an immediate simplified reply or a complete one within fifteen days) and imposes no equivalent uniform one-month bound, so the numeric threshold used for CQ28 would have to be re-derived from Brazilian law. Second, our reading of "systematic" as automated and pre-arranged rests on WP29 and EDPB guidance (12), which are organs of the GDPR framework; the Brazilian authority (ANPD) has issued no equivalent interpretation, so that parameter for CQ51 would need local grounding or would remain an open interpretive choice. Since these values are exposed as explicit parameters rather than embedded in the ontology, adapting them is a re-parameterization rather than a re-modeling, but it is a necessary step whose outcome is not guaranteed to match ours.

A final difference is methodological rather than legal: our competency questions are drawn from the Irish DPC checklist (36), a GDPR-specific instrument. A faithful Brazilian counterpart would re-derive the question set from an official LGPD instrument, such as the ANPD guidance for small-scale processing agents (14) already surfaced in Section 3.2, whose questions may not coincide exactly with those we used. Establishing generalization to Brazil therefore amounts to more than re-running our queries over Brazilian data: it means re-parameterizing the interpretive values against the LGPD and re-validating coverage against a Brazilian checklist. We regard this as a well-scoped extension of the present work rather than a threat to its internal validity, and a natural target for the multi-jurisdiction validation outlined above.

Scalability Concerns. While our SPARQL queries successfully operate on the SAPOS dataset instantiation, we have not conducted performance benchmarking for larger-scale enterprise datasets. This limitation is shared across the field: (65) acknowledge the need for future benchmarking of their system components, while (23) present only an initial performance evaluation of their compliance checking algorithm. (85) explicitly notes that their evaluation focuses on feature existence rather than performance metrics or scalability considerations.

The computational complexity of reasoning over temporal constraints and causal relationships, particularly when combined with transitive closure operations in our SPARQL queries (e.g., `rdf:type/rdfs:subClassOf*`), may present challenges in real-world applications with millions of personal data entities. Future work should include systematic performance evaluation using synthetic datasets of varying scales to establish empirical bounds

on query response times and identify optimization opportunities.

Manual Modeling Effort. Covering each new competency question requires a manual qualitative analysis and the introduction of new ontological elements. As quantified in Section 5.5 and discussed in Section 4.2, this cost is bounded and front-loaded: the same fixed pipeline is reused for every question, and most introduced elements specialize established community vocabularies, so the effort does not grow unboundedly with coverage. The unresolved difficulty lies elsewhere. The extraction of model-relevant terms from legal prose still demands expert legal reading, and this step, not query execution, is the genuine scalability bottleneck of the *methodology*. A concrete path forward is a human-in-the-loop pipeline in which language models propose candidate terms and resource alignments from the regulatory text while the modeler retains decision authority: a focused instance of the hybrid architecture discussed below, whose immediate payoff would be to lower the marginal cost of each additional question.

Legal Interpretation Boundaries. Our approach inherits the fundamental limitation of all semantic approaches to legal compliance: the necessity of translating legal language into formal representations. As (15) observe, this constitutes a fundamental linguistic challenge rooted in the divergence between legal and technical vocabularies. While legal practitioners construct terminology grounded in natural language to articulate regulatory requirements, software engineers develop formal symbolic systems governed by computational logic and grammar. Bridging these distinct knowledge domains remains inherently imperfect, as no definitive methodology exists to guarantee equivalent semantic mappings across these vocabularies, complicating efforts to objectively assess the fidelity of different formalizations. However, recent developments with Agentic AI may help solve this gap.

Although we employed official check-lists, reducing part of the subjectivity from the ontology engineering process, some level of interpretation remains present. Terms such as “without delay” and “systematic” require interpretation, and our chosen interpretations, while justified through regulatory guidance and literature analysis, may not align with all legal jurisdictions or future court decisions. We tried to be meticulous in our analysis, however this interpretative layer is inherent to any formalization of legal requirements and requires ongoing validation against evolving legal precedents. In this study, the author assumed the role of the responsible expert, grounding every interpretation in official regulatory sources (the GDPR itself, WP29/EDPB guidance, and the Irish DPC checklist) and explicitly exposing each interpretive choice as a replaceable parameter (Section 4.2). In a real deployment,

that responsibility should sit with a privacy professional – typically the controller's data protection officer or legal counsel – who would also be the natural candidate to perform the task-based validation with domain experts that the NeOn methodology recommends (Section 6).

Provenance Limitations. As highlighted in previous literature (28; 112), deletion and anonymization procedures present inherent challenges for provenance-based models. This limitation stems from a fundamental tension: provenance aims to maintain comprehensive historical records of data, while deletion and anonymization rights require modifying or erasing that very history. While our model captures the execution of these procedures through invalidation relationships, the complete removal of data entities from the provenance graph could compromise the ability to demonstrate compliance with the procedures themselves. Further investigation is needed to balance these competing requirements while maintaining both accountability and data subject rights.

Completeness of the Provenance Record. Our treatment of CQ51 (Section 4.6) carries an assumption worth stating explicitly as a threat to validity: that the provenance graph completely and faithfully records the personal data the controller actually retains. Because that question is answered by an inverse query that looks for data still held past its lawful period, the approach soundly *detects* non-compliance, since any returned record is a genuine violation, but it can only *prove* compliance by reading an empty result under a closed-world assumption. Where the record is incomplete, an empty result may reflect unrecorded data rather than genuine erasure, so the compliance claim is only as strong as the record's completeness. In the GDPR setting this assumption is partly discharged by the controller's Article 30 duty to keep records of processing; nonetheless, guaranteeing that the captured provenance mirrors the live data store falls outside the model and remains a prerequisite for trusting any compliance verdict it supports.

Hybrid Symbolic-Neural Approaches. Our work employs symbolic semantic approaches (77), based on formal ontologies and logical reasoning, by providing explicit, traceable, and verifiable checking mechanisms. However, as (30) observe, recent advances in Large Language Models (LLMs) (including transformer-based architectures such as BERT (41), GPT series (26; 9), and their successors) present opportunities for hybrid approaches that combine the strengths of both paradigms.

While symbolic models excel at formal reasoning and provide transparent decision paths

(critical requirements for regulatory compliance) LLMs could assist in labor-intensive tasks such as extracting compliance requirements from natural language privacy policies, suggesting ontological alignments, or identifying potential gaps in coverage. Future work could explore carefully designed hybrid architectures where LLMs serve as assistive tools for ontology population and policy translation, while symbolic reasoning maintains primacy for compliance verification, thus preserving the traceability and formal guarantees essential for GDPR compliance demonstration.

References

[1] Eur-lex - 32016r0679 - en - eur-lex. https://eur-lex.europa.eu/eli/reg/2016/679/oj. (Accessed on 07/09/2021).

[2] Gdpr enforcement tracker - list of gdpr fines. https://www.enforcementtracker.com/?insights. (Accessed on 03/12/2022).

[3] Prov-overview. https://www.w3.org/TR/2013/NOTE-prov-overview-20130430/. (Accessed on 05/06/2022).

[4] The special usage policy language. https://ai.wu.ac.at/policies/policylanguage/. (Accessed on 09/02/2022).

[5] What is the lgpd? brazil's version of the gdpr - gdpr.eu. https://gdpr.eu/gdpr-vs-lgpd/. (Accessed on 07/09/2021).

[6] Bgbl. teil i: Nr. 7 (1977) - offenegesetze.de – freier zugang zu unseren gesetzen. https://offenegesetze.de/veroeffentlichung/bgbl1/1977/7#page=1, 1978. (Accessed on 12/09/2022).

[7] Bundesverfassungsgericht - entscheidungen - zur verfassungsmäßigkeit des volkszählungsgesetzes 1983. https://www.bundesverfassungsgericht.de/SharedDocs/Entscheidungen/DE/1983/12/rs19831215_1bvr020983.html, december 1983. (Accessed on 12/09/2022).

[8] D2.9 amass platform validation. Tech. rep., AMASS Project Consortium, June 2020. Deliverable D2.9 of the AMASS Project (Architecture-driven, Multi-concern and Seamless Assurance and Certification of Cyber-Physical Systems).

[9] ACHIAM, J.; ADLER, S.; AGARWAL, S.; AHMAD, L.; AKKAYA, I.; ALEMAN, F. L.; ALMEIDA, D.; ALTENSCHMIDT, J.; ALTMAN, S.; ANADKAT, S., ET AL. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[10] AGENCIA ESPAÑOLA DE PROTECCIÓN DE DATOS (AEPD). Listado de cumplimiento normativo del RGPD. https://www.aepd.es/guias/guia-listado-de-cumplimiento-del-rgpd.pdf, 2018. Accessed 2026-05-19.

[11] ALBERTONI, R.; BROWNING, D.; COX, S.; PEREGO, A.; WINSTANLEY, P., ET AL. Data catalog vocabulary (dcat)-version 2, w3c recommendation.

[12] ARTICLE 29 DATA PROTECTION WORKING PARTY. Guidelines on data protection officers (dpos). https://ec.europa.eu/newsroom/article29/items/612048, 2017. (Accessed on 01/24/2026).

[13] ATKINSON, C.; GUTHEIL, M.; KIKO, K. On the relationship of ontologies and models. In *Meta-modelling and ontologies–Proceedings of the 2nd Workshop on Meta-Modelling–WoMM 2006* (2006), Gesellschaft für Informatik e. V.

[14] AUTORIDADE NACIONAL DE PROTEÇÃO DE DADOS (ANPD). Guia orientativo: Segurança da informação para agentes de tratamento de pequeno porte, versão 1.0. https://www.gov.br/anpd/pt-br/centrais-de-conteudo/materiais-educativos-e-publicacoes/guia-orientativo-sobre-seguranca-da-informacao-para-agentes-de-tratamento-de-pequeno-porte, 2021. Accessed 2026-05-19.

[15] AYALA-RIVERA, V.; PASQUALE, L. The grace period has ended: An approach to operationalize gdpr requirements. In *2018 IEEE 26th International Requirements Engineering Conference (RE)* (2018), IEEE, pp. 136–146.

[16] BARTOLINI, C.; MUTHURI, R.; SANTOS, C. Using ontologies to model data protection requirements in workflows. In *JSAI International Symposium on Artificial Intelligence* (2015), Springer, pp. 233–248.

[17] BAYERISCHES LANDESAMT FÜR DATENSCHUTZAUFSICHT (BAYLDA). Datenschutz-checkliste für kleine unternehmen. https://www.lda.bayern.de/media/checkliste/baylda_checkliste_handel_solo.pdf, [n.d.]. Accessed 2026-05-19.

[18] BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. The semantic web. *Scientific american 284*, 5 (2001), 34–43.

[19] BESIK, S. I.; FREYTAG, J.-C. Ontology-based privacy compliance checking for clinical workflows. In *LWDA* (2019), pp. 218–229.

[20] BESIK, S. I.; FREYTAG, J.-C. Managing consent in workflows under gdpr. In *ZEUS* (2020), pp. 18–25.

[21] BOECKL, K.; FAGAN, M.; FISHER, W.; LEFKOVITZ, N.; MEGAS, K. N.; NADEAU, E. M.; PICCARRETA, B. M.; O'ROURKE, D. G.; SCARFONE, K. A. Considerations for managing internet of things (iot) cybersecurity and privacy risks. NIST Interagency/Internal Report (NISTIR) 8228, National Institute of Standards and Technology, June 2019.

[22] BONATTI, P. A.; BOS, B.; DECKER, S.; FERNANDEZ GARCIA, J. D.; KIRRANE, S.; PERISTERAS, V.; POLLERES, A.; WENNING, R. Data privacy vocabularies and controls: Semantic web for transparency and privacy. CEUR Workshop Proceedings.

[23] BONATTI, P. A.; KIRRANE, S.; PETROVA, I. M.; SAURO, L. Machine understandable policies and gdpr compliance checking. *KI-Künstliche Intelligenz 34*, 3 (2020), 303–315.

[24] BRASILEIRO, P. L13709. `http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm`, May 2018. (Accessed on 03/06/2022).

[25] BREAUX, T. D.; ANTÓN, A. I.; SPAFFORD, E. H. A distributed requirements management framework for legal compliance and accountability. *computers & security 28*, 1-2 (2009), 8–17.

[26] BROWN, T. B.; MANN, B.; RYDER, N.; SUBBIAH, M.; KAPLAN, J.; DHARIWAL, P.; NEELAKANTAN, A.; SHYAM, P.; SASTRY, G.; ASKELL, A.; AGARWAL, S.; HERBERT-VOSS, A.; KRUEGER, G.; HENIGHAN, T.; CHILD, R.; RAMESH, A.; ZIEGLER, D. M.; WU, J.; WINTER, C.; HESSE, C.; CHEN, M.; SIGLER, E.; LITWIN, M.; GRAY, S.; CHESS, B.; CLARK, J.; BERNER, C.; MCCANDLISH, S.; RADFORD, A.; SUTSKEVER, I.; AMODEI, D. Language models are few-shot learners. *Advances in Neural Information Processing Systems 33* (2020), 1877–1901.

[27] BURESH, D. L. A comparison between the european and the american approaches to privacy. *Indon. J. Int'l & Comp. L. 6* (2019), 257.

[28] CAMPAGNA, D. P.; DA SILVA, A. S.; BRAGANHOLO, V. Achieving gdpr compliance through provenance: An extended model. In *SBBD* (2020), pp. 13–24.

[29] CAO, Y.; JONES, C.; CUEVAS-VICENTTÍN, V.; JONES, M. B.; LUDÄSCHER, B.; MCPHILLIPS, T.; MISSIER, P.; SCHWALM, C.; SLAUGHTER, P.; VIEGLAIS, D., ET AL. Dataone: a data federation with provenance support. In *International Provenance and Annotation Workshop* (2016), Springer, pp. 230–234.

[30] CARO, J. D.; ROBALDO, L.; WYNER, A. Z. A shacl-based approach for enhancing automated compliance checking with rdf data. *Information 15*, 12 (2024), 759.

[31] CHHETRI, T. R.; KURTEVA, A.; DELONG, R. J.; HILSCHER, R.; KORTE, K.; FENSEL, A. Data protection by design tool for automated gdpr compliance verification based on semantically modeled informed consent. *Sensors 22*, 7 (2022), 2763.

[32] CHIRIGATI, F.; RAMPIN, R.; SHASHA, D.; FREIRE, J. Reprozip: Computational reproducibility with ease. In *Proceedings of the 2016 international conference on management of data* (2016), pp. 2085–2088.

[33] CLIFFORD, B.; FOSTER, I.; VOECKLER, J.-S.; WILDE, M.; ZHAO, Y. Tracking provenance in a virtual data grid. *Concurrency and Computation: Practice and Experience 20*, 5 (2008), 565–575.

[34] COMMISSION NATIONALE DE L'INFORMATIQUE ET DES LIBERTÉS (CNIL). RGPD: par où commencer ? la check-list pour les TPE/PME. `https://www.cnil.fr/sites/cnil/files/atoms/files/check-list_rgpd_pour_les_tpe-pme.pdf`, [n.d.]. Accessed 2026-05-19.

[35] COMMISSIONER, D. P. Statutory reports | data protection commissioner. `https://dataprotection.ie/en/dpc-guidance/publications`. (Accessed on 06/03/2022).

[36] COMMISSIONER, D. P. Guidance for smes | data protection commissioner. `https://www.dataprotection.ie/en/dpc-guidance/guidance-smes`, 2019.

[37] CONSORTIUM, W. W. W., ET AL. Data catalog vocabulary (dcat).

[38] COUNCIL OF EUROPE. Convention for the protection of human rights and fundamental freedoms, 1950.

[39] DALPIAZ, F.; PAJA, E.; GIORGINI, P. Security requirements engineering via commitments. In *2011 1st Workshop on Socio-Technical Aspects in Security and Trust (STAST)* (2011), IEEE, pp. 1–8.

[40] DEBRUYNE, C.; RIGGIO, J.; DE TROYER, O.; O'SULLIVAN, D. An ontology for representing and annotating data flows to facilitate compliance verification. In *2019 13th International Conference on Research Challenges in Information Science (RCIS)* (2019), IEEE, pp. 1–6.

[41] DEVLIN, J.; CHANG, M.-W.; LEE, K.; TOUTANOVA, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2019). Published at NAACL 2019.

[42] DITTEL, A. Data security requirements under gdpr - lexology. https://www.lexology.com/library/detail.aspx?g=1426e18d-f687-45a0-b779-4aeb362a03ac, May 2016. (Accessed on 01/04/2023).

[43] ECHENIM, K. U.; JOSHI, K. P. Iot-reg: A comprehensive knowledge graph for real-time iot data privacy compliance. In *2023 IEEE International Conference on Big Data (BigData)* (2023), IEEE, pp. 2897–2906.

[44] (EDPB), E. D. P. B. "guidelines on the accreditation of certification bodies for the purpose of certification of data controllers and data processors under the general data protection regulation (gdpr)" (wp 260). https://ec.europa.eu/info/publications/guidelines-accreditation-certification-bodies-purpose-certification-data-controllers-and-data-processors-under-general-data-protection-regulation-gdpr-wp260_en. (Accessed on 01/05/2023).

[45] ERICKSON, A. Comparative analysis of the eu's gdpr and brazil's lgpd: Enforcement challenges with the lgpd. *Brook. J. Int'l L. 44* (2018), 859.

[46] EU, G. Gdpr small business survey. *GDPR. EU* (2019).

[47] EUROPEAN DATA PROTECTION BOARD. Guidelines 4/2019 on Article 25 Data Protection by Design and by Default. https://edpb.europa.eu/, 2019. Version adopted after public consultation.

[48] FATEMA, K.; HADZISELIMOVIC, E.; PANDIT, H. J.; DEBRUYNE, C.; LEWIS, D.; O'SULLIVAN, D. Compliance through informed consent: Semantic based consent permission and data management model. *PrivOn@ ISWC 1951* (2017), 1–16.

[49] FREIRE, J.; KOOP, D.; SANTOS, E.; SILVA, C. T. Provenance for computational tasks: A survey. *Computing in Science & Engineering 10*, 3 (2008), 11–21.

[50] GARIJO, D.; GIL, Y. The p-plan ontology. http://vocab.linkeddata.es/p-plan/index.html. (Accessed on 06/03/2022).

[51] GENESERETH, M. R.; NILSSON, N. J. *Logical foundations of artificial intelligence*. Morgan Kaufmann, 2012.

[52] GERL, A.; POHL, D. Critical analysis of lpl according to articles 12-14 of the gdpr. In *Proceedings of the 13th International Conference on Availability, Reliability and Security* (2018), pp. 1–9.

[53] GHARIB, M. Toward an architecture to improve privacy and informational self-determination through informed consent. *Information & Computer Security 30*, 4 (2022), 549–561.

[54] GJERMUNDRØD, H.; DIONYSIOU, I.; COSTA, K. privacytracker: a privacy-by-design gdpr-compliant framework with verifiable data traceability controls. In *International Conference on Web Engineering* (2016), Springer, pp. 3–15.

[55] GREENLEAF, G. Global data privacy laws 2019: 132 national laws & many bills.

[56] GREENLEAF, G. Global data privacy laws 2021: Despite covid delays, 145 laws show gdpr dominance.

[57] GREENLEAF, G. Global data privacy laws 2025: 172 countries, twelve new in 2023/24. *Twelve New in 24* (2025).

[58] GREENLEAF, G.; COTTIER, B. Data privacy laws and bills: Growth in africa, gdpr influence. *GDPR Influence (April 12, 2018) 152* (2018), 11–13.

[59] GRUBER, T. R. A translation approach to portable ontology specifications. *Knowledge acquisition 5*, 2 (1993), 199–220.

[60] GUPTA, S. D.; HAHMANN, T. OPPO: An Ontology for Describing Fine-Grained Data Practices in Privacy Policies of Online Social Networks. In *Proceedings of the 9th Joint Ontology Workshops (JOWO 2023)* (2023). Held at the 15th International Conference on Formal Ontology in Information Systems (FOIS 2023), Sherbrooke, Canada.

[61] HLOMANI, H.; STACEY, D. Approaches, methods, metrics, measures, and subjectivity in ontology evaluation: A survey. *Semantic Web Journal 1*, 5 (2014), 1–11.

[62] (ICO), I. C. O. Guide to data protection | ico. `https://ico.org.uk/for-organisations/guide-to-data-protection/`. (Accessed on 01/05/2023).

[63] INFORMATION COMMISSIONER'S OFFICE (ICO). Data protection self assessment. `https://ico.org.uk/for-organisations/advice-for-small-organisations/checklists/data-protection-self-assessment/`, [n.d.]. Accessed 2026-05-19.

[64] J. PANDIT, H.; ESTEVES, B.; P. KROG, G.; RYAN, P.; GOLPAYEGANI, D.; FLAKE, J. Data privacy vocabulary (dpv)–version 2.0. In *International Semantic Web Conference* (2024), Springer, pp. 171–193.

[65] KIRRANE, S.; FERNÁNDEZ, J. D.; DULLAERT, W.; MILOSEVIC, U.; POLLERES, A.; BONATTI, P. A.; WENNING, R.; DROZD, O.; RASCHKE, P. A scalable consent, transparency and compliance architecture. In *European Semantic Web Conference* (2018), Springer, pp. 131–136.

[66] KOOPS, B.-J.; LEENES, R. Privacy regulation cannot be hardcoded. a critical comment on the 'privacy by design'provision in data-protection law. *International Review of Law, Computers & Technology 28*, 2 (2014), 159–171.

[67] KUNER, C. The european commission's proposed data protection regulation: A copernican revolution in european data protection law. *Bloomberg BNA Privacy and Security Law Report (2012) February 6*, 2012 (2012), 1–15.

[68] KURTEVA, A. Making sense of consent with knowledge graphs, 2022.

[69] LAMARI, S.; BENBLIDIA, N.; TIBERMACINE, C.; URTADO, C.; VAUTTIER, S. A process for assisting privacy-by-design software engineering. In *ICSR 2022-20th International Conference on Software and Systems Reuse* (2022).

[70] MAGALIE D. LE CLERC, P. M. Gdpr survey - benefits beyond compliance. `https://www.bakermckenzie.com/-/media/files/insight/publications/2020/04/gdpr_survey.pdf?la=en`, 2020. (Accessed on 06/16/2022).

[71] MATULEVIČIUS, R.; TOM, J.; KALA, K.; SING, E. A method for managing gdpr compliance in business processes. In *International Conference on Advanced Information Systems Engineering* (2020), Springer, pp. 100–112.

[72] MAYRING, P. Qualitative content analysis: Demarcation, varieties, developments. In *Forum: Qualitative social research* (2019), vol. 20, Freie Universität Berlin.

[73] MONTEIRO, R. L. The new brazilian general data protection law—a detailed analysis. *International Association of Privacy Professionals. Retrieved February 4* (2018), 2019.

[74] NADEAU, M. What is the gdpr, its requirements and facts? | cso online. `https://www.csoonline.com/article/3202771/general-data-protection-regulation-gdpr-requirements-deadlines-and-facts.html`, June 2020. (Accessed on 01/04/2023).

[75] NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY (NIST). NIST privacy framework: A tool for improving privacy through enterprise risk management, version 1.0. NIST cybersecurity white paper, National Institute of Standards and Technology, 2020. Accessed 2026-05-19.

[76] NEGRI-RIBALTA, C.; NOEL, R.; HERBAUT, N.; PASTOR, O.; SALINESI, C. Socio-technical modelling for gdpr principles: an extension for the sts-ml. In *2022 IEEE 30th international requirements engineering conference workshops (REW)* (2022), IEEE, pp. 238–243.

[77] NEWELL, A.; SIMON, H. A. Computer science as empirical inquiry: Symbols and search. *Communications of the ACM 19*, 3 (1976), 113–126. Turing Award Lecture.

[78] NOY, N. F.; MCGUINNESS, D. L., ET AL. Ontology development 101: A guide to creating your first ontology, 2001.

[79] O'LEARY, D. E. Review: Ontologies: A silver bullet for knowledge management and electronic commerce. *Comput. J. 48*, 4 (2005), 498.

[80] OLTRAMARI, A.; PIRAVIPERUMAL, D.; SCHAUB, F.; WILSON, S.; SADEH, N.; REIDENBERG, J. Privonto: A semantic framework for the analysis of privacy policies. *Semantic Web 9* (2018), 185–203.

[81] OTTO, P. N.; ANTÓN, A. I. Addressing legal requirements in requirements engineering. In *15th IEEE international requirements engineering conference (RE 2007)* (2007), IEEE, pp. 5–14.

[82] PAJA, E.; DALPIAZ, F.; GIORGINI, P. Modelling and reasoning about security requirements in socio-technical systems. *Data & Knowledge Engineering 98* (2015), 123–143.

[83] PALMIRANI, M.; MARTONI, M.; ROSSI, A.; BARTOLINI, C.; ROBALDO, L. Legal ontology for modelling gdpr concepts and norms. In *Legal Knowledge and Information Systems*. IOS Press, 2018, pp. 91–100.

[84] PALMIRANI, M.; MARTONI, M.; ROSSI, A.; BARTOLINI, C.; ROBALDO, L. Pronto: Privacy ontology for legal compliance. In *Proc. 18th Eur. Conf. Digital Government (ECDG)* (2018), pp. 142–151.

[85] PANDIT, H. J. *Representing Activities associated with Processing of Personal Data and Consent using Semantic Web for GDPR Compliance*. Tese de Doutorado, Trinity College Dublin, 2020.

[86] PANDIT, H. J. Representing activities associated with processing of personal data and consent using semantic web for gdpr compliance. *Trinity College Dublin, School of Computer Science & Statistics* (2020).

[87] PANDIT, H. J.; DEBRUYNE, C.; O'SULLIVAN, D.; LEWIS, D. Gconsent-a consent ontology based on the gdpr. In *European Semantic Web Conference* (2019), Springer, pp. 270–282.

[88] PANDIT, H. J.; ESTEVES, B.; KROG, G. P.; RYAN, P.; GOLPAYEGANI, D.; FLAKE, J. Data privacy vocabulary (dpv)–version 2. *arXiv preprint arXiv:2404.13426* (2024).

[89] PANDIT, H. J.; FATEMA, K.; O'SULLIVAN, D.; LEWIS, D. Gdprtext-gdpr as a linked data resource. In *European Semantic Web Conference* (2018), Springer, pp. 481–495.

[90] PANDIT, H. J.; LEWIS, D. Modelling provenance for gdpr compliance using linked open data vocabularies. In *PrivOn@ ISWC* (2017).

[91] PANDIT, H. J.; O'SULLIVAN, D.; LEWIS, D. Exploring gdpr compliance over provenance graphs using shacl. In *SEMANTICS Posters&Demos* (2018).

[92] PANDIT, H. J.; O'SULLIVAN, D.; LEWIS, D. An ontology design pattern for describing personal data in privacy policies. In *WOP@ ISWC* (2018), pp. 29–39.

[93] PANDIT, H. J.; O'SULLIVAN, D.; LEWIS, D. Towards knowledge-based systems for gdpr compliance. In *CKGSemStats@ ISWC* (2018).

[94] PANDIT, H. J.; O'SULLIVAN, D.; LEWIS, D. Extracting provenance metadata from privacy policies. In *International Provenance and Annotation Workshop* (2018), Springer, pp. 262–265.

[95] PANDIT, H. J.; O'SULLIVAN, D.; LEWIS, D. 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* (2018), vol. 4, p. 5.

[96] PANDIT, H. J.; O'SULLIVAN, D.; LEWIS, D. Queryable provenance metadata for gdpr compliance. *Procedia Computer Science 137* (2018), 262–268.

[97] PARLIAMENT, E.; OF THE EUROPEAN COUNCIL. Eur-lex - 31995l0046. `https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A31995L0046`, October 1995. (Accessed on 12/26/2022).

[98] PETERSEN, K.; FELDT, R.; MUJTABA, S.; MATTSSON, M. Systematic mapping studies in software engineering. In *12th International Conference on Evaluation and Assessment in Software Engineering (EASE) 12* (2008), pp. 1–10.

[99] PIMENTEL, J. F.; FREIRE, J.; BRAGANHOLO, V.; MURTA, L. Tracking and analyzing the evolution of provenance from scripts. In *International Provenance and Annotation Workshop* (2016), Springer, pp. 16–28.

[100] POVEDA-VILLALÓN, M.; GÓMEZ-PÉREZ, A.; SUÁREZ-FIGUEROA, M. C. Oops!(ontology pitfall scanner!): An on-line tool for ontology evaluation. *International Journal on Semantic Web and Information Systems (IJSWIS) 10*, 2 (2014), 7–34.

[101] RAAD, J.; CRUZ, C. A survey on ontology evaluation methods. In *International conference on knowledge engineering and ontology development* (2015), vol. 2, SciTePress, pp. 179–186.

[102] ROBOL, M.; SALNITRI, M.; GIORGINI, P. Toward gdpr-compliant socio-technical systems: Modeling language and reasoning framework. In *The Practice of Enterprise Modeling: 10th IFIP WG 8.1. Working Conference, PoEM 2017, Leuven, Belgium, November 22-24, 2017, Proceedings 10* (2017), Springer, pp. 236–250.

[103] RUIZ, A.; MARTIN, Y.-S.; MARTINEZ, J.; QUINTANS, J.; MOCKLY, G.; GYRARD, A.; CREPAX, T. Modeling ecosystems of reference frameworks for assurance: a case on privacy impact assessment regulation and guidelines. *Software and Systems Modeling 22*, 4 (2023), 1175–1196.

[104] RYAN, P.; BRENNAN, R. Support for enhanced gdpr accountability with the common semantic model for ropa (csm-ropa). *SN Computer Science 3*, 3 (2022), 224.

[105] RYAN, P.; BRENNAN, R.; PANDIT, H. J. Dpcat: Specification for an interoperable and machine-readable data processing catalogue based on gdpr. *Information 13*, 5 (2022), 244.

[106] RYAN, P.; PANDIT, H. J.; BRENNAN, R. Building a data processing activities catalog: Representing heterogeneous compliance-related information for gdpr using dcat-ap and dpv. In *SEMANTiCS* (2021), pp. 169–182.

[107] SUÁREZ-FIGUEROA, M. C.; GÓMEZ-PÉREZ, A.; FERNÁNDEZ-LÓPEZ, M. The neon methodology for ontology engineering. In *Ontology engineering in a networked world*. Springer, 2012, pp. 9–34.

[108] SUÁREZ-FIGUEROA, M. C.; GÓMEZ-PÉREZ, A.; MOTTA, E.; GANGEMI, A. *Ontology engineering in a networked world*. Springer, 2011.

[109] TERUEL, M.; CARDELLINO, C.; CARDELLINO, F.; ALEMANY, L. A.; VILLATA, S. Legal text processing within the mirel project. In *1st Workshop on Language Resources and Technologies for the Legal Knowledge Graph* (2018), p. 42.

[110] TOM, J.; SING, E.; MATULEVIČIUS, R. Conceptual representation of the gdpr: model and application directions. In *International Conference on Business Informatics Research* (2018), Springer, pp. 18–28.

[111] TORRE, D.; ALFEREZ, M.; SOLTANA, G.; SABETZADEH, M.; BRIAND, L. Modeling data protection and privacy: application and experience with gdpr. *Software and Systems Modeling 20*, 6 (2021), 2071–2087.

[112] UJCICH, B. E.; BATES, A.; SANDERS, W. H. A provenance model for the european union general data protection regulation. In *International Provenance and Annotation Workshop* (2018), Springer, pp. 45–57.

[113] WOHLIN, C. Guidelines for snowballing in systematic literature studies and a replication in software engineering. In *Proceedings of the 18th international conference on evaluation and assessment in software engineering* (2014), pp. 1–10.

APPENDIX A – Competency Questions

A.1

Table A.1 lists all Competency Questions defined by the (36).

Table A.1: Competency Questions published by (36) used as Compliance Question for this paper

Code	Description
CQ01	List the categories of data subjects and personal data collected and retained e.g. current employee data; retired employee data; customer data (sales information); marketing database; CCTV footage.
CQ02	List each type of personal data included within each category of personal data e.g. name, address, banking details, purchasing history, online browsing history, video and images.
CQ03	List the source(s) of the personal data e.g. collected directly from individuals; from third parties (if third party identify the data controller as this information will be necessary to meet obligations under Article 14).
CQ04	Within each category of personal data list the purposes for the data is collected and retained e.g. marketing, service enhancement, research, product development, systems integrity, HR matters, advertising.
CQ05	For each purpose that personal data is processed, list the legal basis on which it is based e.g. consent, contract, legal obligation (Article 6).
CQ06	If special categories of personal data are collected and retained, set out details of the nature of the data e.g. health, genetic, biometric data.
CQ07	List the legal basis on which special categories of personal data are collected and retained e.g. explicit consent, legislative basis (Article 9).

Continued on next page

Table A.1 continued from previous page

Code	Description
CQ08	"or each category of personal data, list the period for which the data will be retained e.g. one month? one year? As a general rule data must be retained for no longer than is necessary for the purpose for which it was collected in the first place."
CQ09	Identify actions that are required to ensure all personal data processing operations are GDPR compliant e.g. this may include deleting data where there is no further purpose for retention.
CQ10	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?
CQ11	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?
CQ12	Are procedures in place to demonstrate that an individual has consented to their data being processed?
CQ13	Are procedures in place to allow an individual to withdraw their consent to the processing of their personal data?
CQ14	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?
CQ15	"If legitimate interest is a legal basis on which personal data is processed, has an appropriate analysis been carried out to ensure that the use of this legal basis is appropriate? (That analysis must demonstrate that 1) there is a valid legitimate interest, 2) the data processing is strictly necessary in pursuit of the legitimate interest, and 1) the processing is not prejudicial to or overridden by the rights of the individual)"
CQ16	Is there a documented policy/procedure for handling Subject Access Requests (SARs)?
CQ17	Is your organisation able to respond to SARs within one month?
CQ18	Are procedures in place to provide individuals with their personal data in a structured, commonly used and machine readable format?

Continued on next page

Table A.1 continued from previous page

Code	Description
CQ19	Are there controls and procedures in place to allow personal data to be deleted or rectified (where applicable)?
CQ20	Are there controls and procedures in place to halt the processing of personal data where an individual has on valid grounds sought the restriction of processing?
CQ21	Are individuals told about their right to object to certain types of processing such as direct marketing or where the legal basis of the processing is legitimate interests or necessary for a task carried out in the public interest?
CQ22	Are there controls and procedures in place to halt the processing of personal data where an individual has objected to the processing?
CQ23	If automated decision making, which has a legal or significant similar affect for an individual, is based on consent, has explicit consent been collected?
CQ24	Where an automated decision is made which is necessary for entering into, or performance of, a contract, or based on the explicit consent of an individual, are procedures in place to facilitate an individual's right to obtain human intervention and to contest the decision?
CQ25	Have the circumstances been documented in which an individual's data protection rights may be lawfully restricted? Note: the Irish Data Protection Bill will set out further details on the implementation of Article 23.
CQ26	Are personal data only used for the purposes for which they were originally collected?
CQ27	Are the personal data collected limited to what is necessary for the purposes for which they are processed?
CQ28	Are procedures in place to ensure personal data are kept up to date and accurate and where a correction is required, the necessary changes are made without delay?
CQ29	Are retention policies and procedures in place to ensure data are held for no longer than is necessary for the purposes for which they were collected?
CQ30	Is your business subject to other rules that require a minimum retention period (e.g. medical records/tax records)?
CQ31	Do you have procedures in place to ensure data is destroyed securely, in accordance with your retention policies?

Continued on next page

Table A.1 continued from previous page

Code	Description
CQ32	Are procedures in place to ensure that there is no unnecessary or unregulated duplication of records?
CQ33	Are service users/employees fully informed of how you use their data in a concise, transparent, intelligible and easily accessible form using clear and plain language?
CQ34	Where personal data are collected directly from the individuals, are procedures in place to provide the information listed at Article 13 of the GDPR?
CQ35	Are procedures in place to ensure personal data are kept up to date and accurate and where a correction is required, the necessary changes are made without delay?
CQ36	If personal data is not collected from the subject but from a third party (e.g. acquired as part of a merger) are procedures in place to provide the information listed at Article 14 of the GDPR?
CQ37	When engaging with individuals, such as when providing a service, sale of a good or CCTV monitoring, are procedures in place to proactively inform individuals of their GDPR rights?
CQ38	Is information on how the organisation facilitates individuals exercising their GDPR rights published in an easily accessible and readable format?
CQ39	Have agreements with suppliers and other third parties processing personal data on your behalf been reviewed to ensure all appropriate data protection requirements are included?
CQ40	Do you need to appoint a DPO as per Article 37 of the GDPR?
CQ41	If it is decided that a DPO is not required, have you documented the reasons why?
CQ42	Where a DPO is appointed, are escalation and reporting lines in place? Are these procedures documented?
CQ43	Have you published the contact details of your DPO to facilitate your customers/ employees in making contact with them?
CQ44	Have you notified your data protection authority (such as the DPC) of your DPO's contact details?

Continued on next page

Table A.1 continued from previous page

Code	Description
CQ45	If your data processing is considered high risk, do you have a process for identifying the need for, and conducting of, DPIAs? Are these procedures documented?
CQ46	Have you assessed the risks involved in processing personal data and put measures in place to mitigate against them?
CQ47	Is there a documented security programme that specifies the technical, administrative and physical safeguards for personal data?
CQ48	Is there a documented process for resolving security related complaints and issues?
CQ49	Is there a designated individual who is responsible for preventing and investigating security breaches?
CQ50	Are industry standard encryption technologies employed for transferring, storing, and receiving individuals' sensitive personal information?
CQ51	Are personal data systematically destroyed, erased, or anonymised when they are no longer legally required to be retained.
CQ52	Can access to personal data be restored in a timely manner in the event of a physical or technical incident?
CQ53	Does the organisation have a documented privacy and security incident response plan?
CQ54	Are there procedures in place to notify the office of the Data Protection Commissioner of a data breach?
CQ55	Are there procedures in place to notify data subjects of a data breach (where applicable)?
CQ56	Are plans and procedures regularly reviewed?
CQ57	Are all data breaches fully documented?
CQ58	Are there cooperation procedures in place between data controllers, suppliers and other partners to deal with data breaches?
CQ59	Is personal data transferred outside the EEA, e.g. to the US or other countries?
CQ60	Does this include any special categories of personal data?
CQ61	What is the purpose(s) of the transfer?
CQ62	Who is the transfer to?

Continued on next page

Table A.1 continued from previous page

Code	Description
CQ63	Are all transfers listed - including answers to the previous questions (e.g. the nature of the data, the purpose of the processing, from which country the data is exported and which country receives the data and who the recipient of the transfer is?)
CQ64	Is there a legal basis for the transfer, e.g. EU Commission adequacy decision; standard contractual clauses. Are these bases documented?
CQ65	Are data subjects fully informed about any intended international transfers of their personal data?

APPENDIX B – Published paper in SBBD 2020

Achieving GDPR Compliance through Provenance: An Extended Model

Daniel P. Campagna[1], **Altigran S. da Silva**[2], **Vanessa Braganholo**[1]

[1]Instituto de Computação – Universidade Federal Fluminense (UFF)
Niterói – RJ – Brazil

[2]Instituto de Computação – Universidade Federal do Amazonas (UFAM)
Manaus – AM – Brazil

danielcampagna@id.uff.br, alti@icomp.ufam.edu.br, vanessa@ic.uff.br

Abstract. *The approval of the General Data Protection Regulation (GDPR) brought a revolution in the way we treat data produced in digital media. The GDPR increases individuals' participation in the treatment of their data, and it also introduces technical challenges, whose failure can lead to a fine of 4% of the organization's annual revenue. Among many approaches that aim to contribute to the solutions of challenges introduced by GDPR, there is a research branch promoting the use of data provenance as a means to make transparent the increasingly complex workflows of systems. However, existing provenance models are not fully compliant with the GDPR. In this paper, we aim to contribute to the evolution of the GDPR data provenance model proposed by Ujcich et al.. We suggest eleven new changes that make the model more apparent and more compatible with the GDPR text. We also present two design patterns that should guide us in using these changes in real contexts.*

Resumo. *A aprovação do Regulamento Geral de Proteção de Dado (GDPR) trouxe uma revolução na maneira como tratamos os dados produzidos em meios digitais. A GDPR inclui uma maior participação dos indivíduos no tratamento dos seus dados e também introduz desafios técnicos cuja preterição pode levar a uma multa de 4% da receita anual da organização. Dentre muitas abordagens que buscam contribuir na solução dos desafios introduzidos pela GDPR, existe um ramo que tem promovido o uso de proveniência de dados como um meio para tornar transparente os passos cada vez mais complexos dos sistemas. No entanto, modelos de proveniência existentes não são completamente aderentes à GDPR. Neste artigo, buscamos contribuir com a evolução do modelo de proveniência de dados da GDPR proposto por Ujcich et al.. Ao final, sugerimos onze novas mudanças que tornam o modelo mais claro e mais compatível com o texto da GDPR, além de dois padrões de projeto que nos orientam em como usar essas mudanças em contextos reais.*

1. Introduction

In April 2016, the European Parliament introduced a Copernican revolution in the European Data Privacy law. The famous movement proposed by Immanuel Kant in philosophy, in the 18th century, shifted the manner we build knowledge by moving away from concepts embodied in the external objects towards the judgments that run in

an individual's mind. The General Data Protection Regulation (GDPR) [Council of European Union 2016] brings up "a similar revolution in European data protection law by seeking to shift its focus away from paper-based, bureaucratic requirements and towards compliance in practice, harmonization of the law, and individual empowerment." [Kuner 2012]. In practice, since the GDPR came into force on May 25th, 2018, European citizens (*subjects*) have the right to be informed of legal compliance when their data (*personal data*) are collected, used, stored, and shared. *Subjects* are also called upon to allow, deny, or restrict the use of their data, in such a manner that companies under GDPR Article 6 can only process data after procuring subject's consent.

A 2016 survey [Tankard 2016] has reported a lack of technologies to aid in evaluating whether *personal data* were processed and stored according to the owner's consent. These results, somehow, reflect our current scene. A recent survey [GDPR.EU 2019] from May 2019 – involving 716 small business leaders in Spain, the United Kingdom, France, and Ireland – reported that only 44% of organizations agree that their organization "describes its data processing activities in clear, plain language to data subjects" [GDPR.EU 2019, Art. 12]. In a word, at a 95% confidence level (+/-4% error margin), the survey authors report that millions of small businesses, only in European territories, are still ignoring part of the GDPR requirements. The penalties for organizations that do not comply with GDPR can be up to €20 million or 4% of their annual revenue.

In the literature, provenance mechanisms have been used to aid in meeting these requirements [Ujcich et al. 2018, Aldeco Perez and Moreau 2008, Martin et al. 2012, Bartolini et al. 2015]. According to the *Oxford English Dictionary*, provenance is "the source or origin of an object; its history and pedigree; a record of the ultimate derivation and passage of an item through its various owners." The World Wide Web Consortium (W3C) defines PROV, an extensible provenance model that uses three main concepts: *entities*, *activities*, and *agents* [Moreau and Missier 2013]. The model then proposes seven relations that can be used to link these concepts. The three concepts could be comprehended as vertices of a provenance graph, while the relations are edges that connect these vertices.

From the work that addresses the GDPR problem using provenance [Ujcich et al. 2018, Aldeco Perez and Moreau 2008, Martin et al. 2012, Bartolini et al. 2015], the model proposed by Ujcich et al. [2018] is the only one that is based on the GDPR law – others are based on previous versions of the law, or on country specific laws. Such model encodes GDPR semantics into subclasses of PROV components, using ontologies [Bartolini et al. 2015, Pandit and Lewis 2017]. The authors also propose three use cases leaning on an example used in prior work that involves collecting personal data for a retail shop. Despite all contributions, we found this model is missing an analysis concerning the law principles.

This paper addresses this gap by identifying and reporting limitations we find when comparing the GDPR data provenance model proposed by Ujcich et al. [2018] and some relevant articles of the law. Also, for each limitation, we propose extensions to the original model. Finally, we present a comprehensive practical case study using this new extended model.

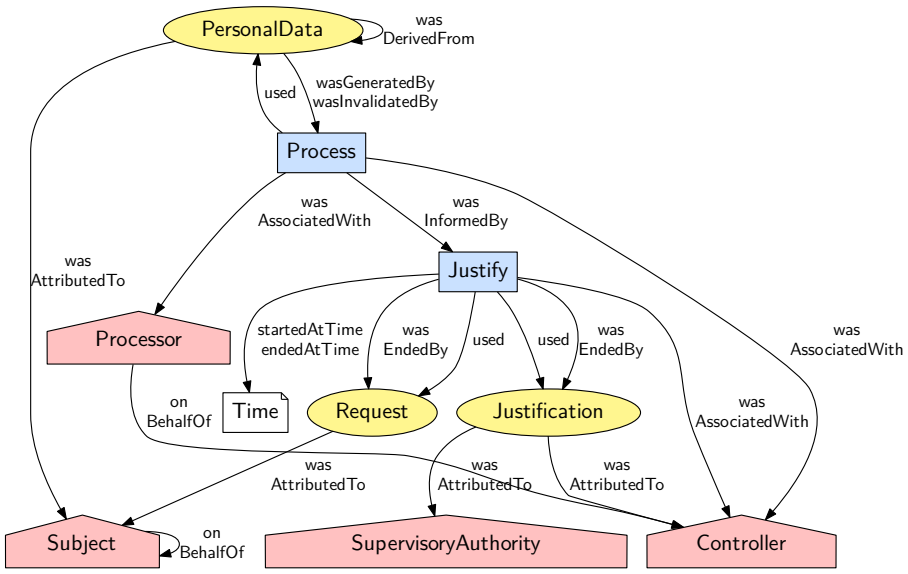

Figure 1. Ujcich et al.'s [2018] GDPR data provenance model with high-level classes.

The paper is organized as follows. Section 2 presents related work regarding the use of data provenance to cope with the GDPR requirements. Section 3 reviews the GDPR data provenance model proposed by Ujcich et al. [2018] . Section 4 discusses the limitations and introduces our extension to the GDPR data provenance model. Section 5 proposes two design patterns involving gathering and maintaining personal data for an on-line retail shop. Section 6 discusses future work and concludes this paper.

2. Related Work

Provenance is often used in e-science contexts to help in tasks related to the interpretation and understanding of results [Freire et al. 2008]. However, this need is also found in research that address problems involving data protection. For instance, Aldeco-Pérez and Moreau [2008] use provenance techniques to aid in auditing the United Kingdom's Data Protection (1998) Act. Martin et al. [2012] describe how provenance can help in tasks such as disclosing, access control, and data usage. Their context involves the earlier Data Protection Directive (1995) [Council of European Union 2016]. Moreover, other researchers explore provenance to solve different problems [Bonatti et al. 2017, Basin et al. 2018, Gjermundrød et al. 2016, Bier 2013].

Regarding the use of data provenance to address GDPR-related challenges, Ujcich et al. [2018] propose a GDPR data provenance model based on the data-processing components proposed in prior work [Bartolini et al. 2015, Pandit and Lewis 2017]. They use a subset of Bartolini et al. [2015] ontology, which represents knowledge about the obligations and rights that agents have among themselves. They also extend the proposal by Pandit and Lewis [2017] that introduces GDPRov. GDPRov extends the P-Plan ontology [Garijo and Gil 2013], which models expected workflow by using PROV's prov:Plan. Ujcich et al. 's model avoids using plans, as it requires pre-specification of the workflow. In exchange, the authors argue that more flexible specifications of how data can be used are possible. In general terms, their main contribution

is a translation of the rough text of the GDPR into a readable provenance language by recycling prior GDPR ontologies to map GDPR concepts into the W3C PROV model.

To deal with the practical issues of collecting provenance, there are whole-system provenance capture mechanisms such as Hi-Fi [Pohly et al. 2012], CamFlow [Pasquier et al. 2015], and Linux Provenance Module [Bates et al. 2015]. These mechanisms capture provenance information at the kernel-level for Linux-based operating systems. However, these systems are too fine-grained, which makes it difficult to use them to enforce data privacy policies.

There are also approaches not based on provenance data. Shastri et al. [2019] handle three traditional database management systems to ensure GDPR compliance. Pandit et al. [2019] demonstrate through use-cases that semantic representations of processes are useful towards maintaining ongoing GDPR compliance. Wang et al. [2019] propose a new paradigm for storing sensitive data. Their approach consists in encapsulating the personal data within a policy which governs its use. Therefore, to be accessed, a given piece of data needs to be submitted to a declassification step which checks whether its policy has been satisfied.

3. Background

Since this paper aims to analyze Ujcich et al.'s [2018] GDPR data provenance model (GDPR model), in this section we introduce an overview of this model. In their paper, the authors summarize six rights and obligations they identify in the GDPR text. Based on the text and recycling prior models, they propose their provenance model which is shown in Figure 1. This Figure graphically represents the high-level classes and relations of their proposed model. Table 1 contains a short explanation of each class of the model. They also propose several subclasses that enhance the model semantics towards the GDPR concepts.

For the purpose of the case study of this paper, we highlight the subclasses of Process and Justify. The remaining subclasses and a more detailed explanation are available in Ujcich et al.'s paper . Process has eight subclasses: Store, which represents an action in the past when some data were persisted; Disclose, which is useful to represent any data transmission between controllers and processors; Pseudonymize that represents "processing of personal data [so that it] can no longer be attributed to a specific data subject without the use of additional information". [Council of European Union 2016, Art. 4]. The remaining subclass are Collect, Retrieve, Combine, Erase, and Profile. Finally, the GDPR data provenance model expresses that any processing lawfulness should be informed by a *subject*'s Consent, or a *controller*'s Obligation, Interest or Authority [Council of European Union 2016, Art. 6]. In the model, these are all subclasses of Justify.

4. Extension to the GDPR Data Provenance Model

Now that we have the GDPR data provenance model [Ujcich et al. 2018] in place, it is possible to map the GDPR concepts into provenance components. However, while doing that, we found some limitations, for which we introduce extensions. In our efforts to propose extensions to this model, we have made a non-systematic read and interpretation and tried to fit the GDPR data provenance model components into each of

Component	Class	Explanation
Agent	Subject	An "identifiable natural person [...] who can be identified, directly or indirectly, in particular by reference to an identifier."
	Controller	An organization "which [...] determines the purposes and means of the processing of personal data."
	Processor	An organization "which processes personal data on behalf of the controller."
	Supervisory Authority	is "An independent public authority" [Council of European Union 2016, Arts. 4, 51-59] "that can monitor and enforce the application of" the GDPR and "handle complaints lodged by a data subject [...] and investigate." [Council of European Union 2016, Art. 57].
Activity	Process	"Any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means."
	Justify	"The rationale that a controller uses in taking some action on personal data, which includes temporal notions of 'start' and 'end' times."
Entity	PersonalData	An "identifier [of a subject] such as a name, an identification number, location data, an on-line identifier or to one or more factors specific to the [...] identity of that natural person" [Council of European Union 2016, Art. 3].
	Request	A request sent from a Subject to a Controller for lawful processing.
	Justification	A justification for lawful processing.

Table 1. Classes of the PROV components proposed by Ujcich et al. [2018] .

the GDPR articles. In cases where it is not clear how a given GDPR article can fit the model, we discuss and propose extensions.

We found eleven limitations in which the model does not cover a concept. In this section, we present extensions that address all these limitations. However, we only discuss two of the proposed extensions (Sections 4.1 and 4.2) due to space constraints. The remaining are discussed in the project documentation page [1].

Tables 2-4 summarize all the limitations and extensions we propose. We classify all limitations and extensions into three types according to their impact on the model. The first set of extensions, which we call plumbing extensions, affects how the GDPR model works, and appends new elements (Table 2 summarizes them). The remaining are called porcelain (Table 3) and wallpaper (Table 4) extensions. Both intend to clarify and enrich this GDPR data provenance model by either suggesting new subclasses representing the semantic meanings of some part of the GDPR text or appending further information detailed in the law.

4.1. Provide the data subject with the period for which their data will be stored

Our first concern addresses the period for which each *personal data* will be stored. Both articles 13 and 14 [Council of European Union 2016, Arts 13, 14] prescribe which further information the *controllers* must provide to *subjects* whose *personal data* are collected. Into those, we found that the GDPR model does not cover the point a), which forces *controllers* to inform to *subjects* "the period for which [their] personal data will be stored [...]" [Council of European Union 2016, point a) of Article 12(2)]. Currently, the GDPR model represents this period using the startedAtTime and endedAtTime rela-

[1] https://dew-uff.github.io/gdpr-data-provenance-model/

GDPR Text	Limitation	Extension
"Where personal data relating to a data subject are collected from the data subject, the controller shall provide [...] the fact that the controller intends to transfer personal data to a third country or international organization [...]." [Council of European Union 2016, point f) of Art. 13(1)].	The GDPR model enables subjects to know whether the controller intends to transfer their data; however, the model does not provide a direct relationship between the controllers.	(i) A new self-relationship (wasAttributedTo) for Controller; and (ii) use Disclose class to represent this transfer.
"The controller shall [...] provide the data subject with [...] the period for which [her] personal data will be stored, or if that is not possible, the criteria used to determine that period" [Council of European Union 2016, point a) of Art. 13(2)].	Although it is possible to represent when a purpose ends, the authors of the GDPR model design patterns explicitly suggest that the obtention of that purpose-ending information does not occur at the beginning.	When *personal data* are obtained, both startedAtTime and endedAtTime relations should be created with their Time annotation, denoting the *Valid Time* [Ozsoyoglu and Snodgrass 1995], those *personal data* will be stored.
"Where the controller intends to further process [...] for a purpose other than that for which the personal data were collected, the controller shall [...]" reveal to data subject its intentions. [Council of European Union 2016, Art. 13(3)].	The current model does not inform whether a Justify is compatible with another one.	(i) A new self-relationship (wasCompatibleWith) for Justify; (ii) A new PseudonymizedData class, as a PersonalData subclass, which instances representing pseudonymized information from all person; and (iii) all those further processes informed by a *Justify* that wasCompatibleWith another must only use PseudonymizedData [Council of European Union 2016, point e) of Art. 6(4)].

Table 2. Plumbing extensions change the GDPR data provenance model high-level classes or relations by adding new elements or changing their use.

tions, which connects Justify entities to Time annotations, thus preserving the timing information of activity compliance. However, the text of this provenance model design suggests that those relations should be created either when its Justify entity is logically present in the database (*i.e.*, creating startedAtTime relation) or when a specific Request is made (*i.e.*, creating endedAtTime relation). This method to capture timing information is known in the literature by *Transaction Time* [Ozsoyoglu and Snodgrass 1995]. An example that enforces this understanding is the first design pattern introduced by Ujcich et al., which describes the use of this provenance model when a user "registers with and provides [her] personal data [...], along with [her] consent" [Ujcich et al. 2018, p. 5]. We notice that the GDPR model creates only the startedAtTime relation at this moment.

In order to represent the period that *personal data* will be stored, we understand

Limitation	Extension
The further processing for archiving purposes in the public interest, scientific or historical research purposes, or statistical purposes are often cited in the text of the law [Council of European Union 2016, Art. 5, 9, 14, 17, 21, 89].	A new ArchivingResearchingStatisticalPurposesJustify class, as sublclass of Justify, which represents those purposes.
Article 9 lists a set of purposes categories of personal data, which demands special justification to process. [Council of European Union 2016, Art. 9(1)]	We suggest the following new classes, all of them subclasses of PersonalData: RacialEthnicData, PoliticalOpinionsData, BeliefData, MembershipData, GeneticData, BiometricData, HealthData, SexLifeData, and SexualOrientationData.
Points b), d), e), and i) of Article 9 explain situations in which the special categories of personal data must not be prohibited [Council of European Union 2016, Art. 9(2)].	We suggest the following new classes: EmploymentSecurityObligation, as a subclass of LegalObligation (point b)); CommunityInterest (point b)) and PublicHealthInterest (point i)), both subclasses of PublicInterest; and PublicData, as a subclass of PersonalData, and Publish, as a subclass of Process (point e)).
Article 16 lays down rectification processes; however, the GDPR model does not propose a class representing that.	A new Rectification, subclass of Process, that represents a rectification process.
Article 21 defines the right of objection; however, the GDPR model does not propose a class to represent this right.	A new class ObjectionRequest, as a subclass of Request, that represents a subject's objection.

Table 3. Porcelain extensions are focused on improving the GDPR data provenance model understanding by providing new subclasses that better express some GDPR articles' points.

that it is enough to create both startedAtTime and endedAtTime relations, with their Time annotation denoting the *Valid Time* [Ozsoyoglu and Snodgrass 1995]. In a word, at the time the *personal data* are collected, along with its consents, *controllers* must inform to the GDPR data provenance model the real-time period this data could be stored. In terms of this provenance model, it means creating both startedAtTime and endedAtTime relations, at the time of data collection.

4.2. Reveal to data subject further processes

Another limitation we have found in this model is the representation of the data's use for a purpose other than the one for which they were collected. Article 5 establishes that "personal data shall be [...] collected for specified [...] purposes and not further processed in a manner that is incompatible with those purposes" [Council of European Union 2016, point b) of Art. 5(1)]. However, this same article sets out that some special types of further processing are not considered incompatible with those initial purposes, since these processes comply with Article 89. In other words, ensuring the principle of data minimization, special types of treatments can be processed regardless of the controller holding a specific justification. This brings us two challenges: (i) it must be clear how our model should represent these special type of further processing, and (ii) how to ensure that the *subject* is aware of these processes even though these processing use pseudonymized data.

We understand that it is possible to represent these treatments through the was-

GDPR Text	Limitation	Extension
"The data subject shall have the right to obtain from the controller [...] any available information as to their source." [Council of European Union 2016, point g) of Art. 15(1)].	The GDPR model does not present any information about the data source.	A new SourceExplanation entity, in which a PersonalData is associated with it, representing a document that describes the source of this data.
"The data subject shall have the right to obtain from the controller [...] the existence of automated decision-making [...] [and] meaningful information about the logic involved [...]." [Council of European Union 2016, point h) of Art. 15(1)].	The GDPR model does not present any that different information about the existence of automated decision-making.	A new AlgorithmExplanation entity, in which an automated Process is associated with it, representing a document that describes the logic involved in that process.
"Where personal data are transferred [...] the data subject shall have the right to be informed of the appropriate safeguards [...]." [Council of European Union 2016, Art. 15(2)] such as describe Art. 46, 47, and 49, also "the means by which to obtain a copy of them or where they have been made available" [Council of European Union 2016, point f) of Art. 13(1), 14(1)].	The GDPR model does not represent any information about the appropriate safeguards of the controller, by concerning their availability to lawfully transfer of personal data.	A new EvaluationOfAdequacy entity, in which a Controller is associated with it, representing a document that describes that required information.

Table 4. Wallpaper extensions present new additions to enrich the model with the information to be compliant with GDPR, although there are other means to present that information to the data owners.

InformedBy relation with an already given Justify activity. However, we see that this introduces ambiguity to the model, due to the representation of two different consents from a single Justify activity. Thus, we propose the wasCompatibleWith self-relation in the Justify activity. Article 6(4) prescribes the rules for valid compatibility between two Justify activities.

In addition, the *subject* should be aware "prior to that further processing" [Council of European Union 2016, Art. 13(3)]. The problem is that Article 89 imposes pseudonymization, which means that is not possible to derive the owner of the pseudonymized data anymore. In order to represent that limitation, we propose the new PseudonymizedData entity, which can not be *attributed to* a Subject. Each instance of this entity represents a pseudonymized property of all *subjects* that have that pseudonymized property. For example, suppose the *subjects* Alice and John, having the same blood type, have registered these data along with their consent in a clinical system. Later, after the *controller* pseudonymizes the blood-type data of all their clients, the provenance graph uses a single instance of PseudonymizedData to represent all pseudonymized blood-type data, including Alice and John's blood type. Following this solution, Alice (and John) are aware that her data could be in use, even though it is not possible to attribute whether the pseudonymized blood-type data belongs to her or to John. Finally, all those further processes should, now, use these PseudonymizedData entities.

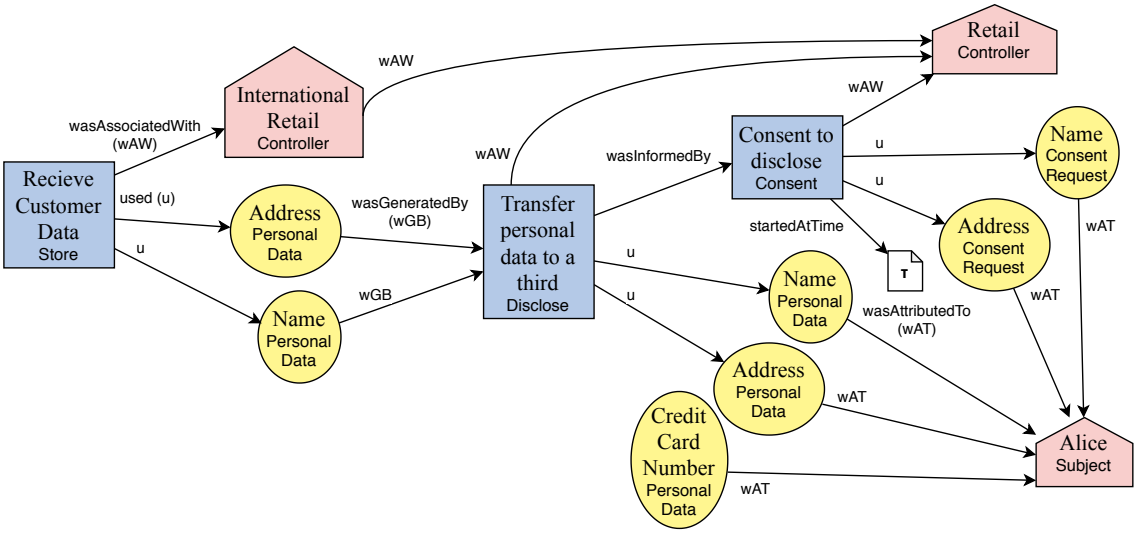

Figure 2. Holding Alice's consent, the retail shop intends to transfer Alice's data to an international retail shop controller. Note that Alice does not consent the controller to share her credit card number with a third party.

5. Design Patterns

Our goal in this section is to present design patterns to clarify *how* the extensions we proposed should be used in practice. Following the GDPR model authors' steps, we use the same running example based on the examples from prior work [Ujcich et al. 2018, Pandit and Lewis 2017, Basin et al. 2018] that involve collecting personal data for an on-line retail shop. Similarly, we assume a customer, Alice, interacts with the retailer by registering, making purchases, and subscribing to marketing information. For this work, we focus only on cases that require our extensions.

5.1. Transfer Data to an International Controller

After time τ, the retail shop intends to transfer Alice's data to an international retail *controller*. Note the association between both *controllers*, which makes it easier to identify the existence of the relation between them through wasAssociatedWith. Figure 2 shows the provenance graph generated from those activities.

Since we inherit the provenance model and did not discard any original component, we still represent a consent for personal data with purposes as a design pattern in the provenance graph by connecting Consent activities to ConsentRequest entities with the used relation. As shown in Figure 2, Alice does not consent to disclose her credit card information to a third party. Note that this use case could not be represented using only the original provenance model [Ujcich et al. 2018].

5.2. Further Process for Other Purposes

At the time τ, Alice registers with and provides her data to the retail shop, along with her consent. At the same time, Alice informs her wish that her data must be erased at the time $\tau + \delta$.

At the time $\tau + 1$, the retail shop intends to calculate the number of customers who share the same city using the address of all customers for statistical purposes.

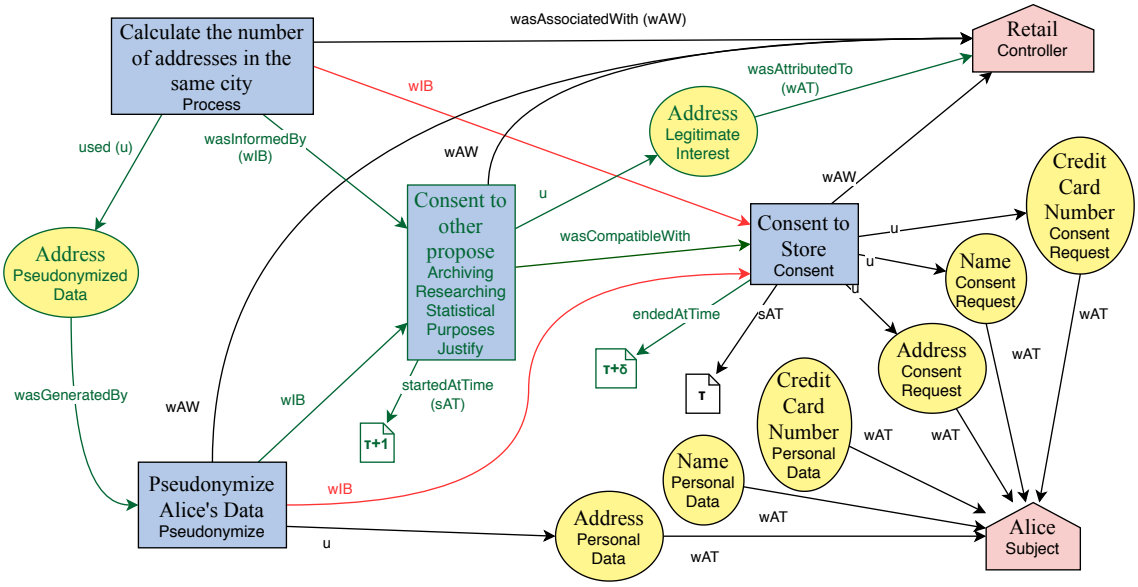

Figure 3. At the time $\tau + 1$, the retail shop intends to calculate the number of addresses in the same city by using personal data. This calculation uses pseudonymized data, and its purpose was not considered incompatible with Alice's consent for storing. Note that, before time $\tau + \delta$, Alice can be informed of the period in which her data will be stored. Note that we represent two graphs in this figure. The first one (black and red arrows) uses the original provenance model, The second one (black and green arrows) uses our extended model.

Thus, the retail shop uses customers' pseudonymized-addresses. We represent those pseudonymized addresses in the provenance graph through the PseudonymizedData entity, as shown in Figure 3. There must be only a single pseudonymized-address entity (regardless of the number of registered customers), representing all customers' pseudonymized-addresses generated by the Pseudonymize processes. This design pattern introduces ambiguity into relationships between data and data owners while enabling the subject to be aware of further processing. The latest design pattern presented in this example is the compatibility relationship between two different purposes. Figure 3 shows, in green and black, that the purpose of calculating the number of customers that share the same city wasCompatibleWith Alice's consent to store her data. In red and black, this Figure represents the use case without the proposed extensions. Note, however, that the provenance graph is more complete using our proposed extensions.

6. Conclusion and Future Work

In this paper, we introduce the problem of enforcing compliance with the GDPR from a data provenance perspective. We start by analyzing the GDPR data provenance model by Ujcich et al. [2018] ; after, we report our analysis comparing this model with some GDPR articles. In that effort, we found eleven points in the law text that are still not suitably represented by this model. We classify these limitations we have found into three types, each one expressing a degree of the impact they affect the model. For each of these limitations, we proposed an extension to the original model that addresses it. Finally, we present two design patterns that represent some of those extensions in the

provenance graph.

Although we have made a high-level effort to evolve this model by analyzing the GDPR text, in future work we also consider practical approaches. The path to helping organizations achieve true compliance should consider the high-level criticism aimed at improving a provenance model conjugated with practical measures that involve multidisciplinary and practical issues. Ujcich et al. [2018] anticipate some of these practical issues. They point out to the necessity to deal with metadata in provenance that also requires GDPR-compliance. They also note other problems, such as the need for inter-controllers audits; the mismatched provenance-granularity levels of; and the ensuring fraud-resistance in provenance collection mechanisms.

Finally, although the GDPR will always require some human activity [Basin et al. 2018], we believe that collaborative efforts can result in a mature solution that will minimize bureaucratic tasks and maximize the transparency of processes.

Acknowledgements. The authors would like to thank CAPES and CNPq for partially supporting this work, and João Felipe Pimentel and Maria Luiza Falci for their reviews and helpful comments.

References

Aldeco Perez, R. and Moreau, L. (2008). Provenance-based auditing of private data use. In *BCS International Academic Conference*.

Bartolini, C., Muthuri, R., and Santos, C. (2015). Using ontologies to model data protection requirements in workflows. In *JSAI International Symposium on Artificial Intelligence*, pages 233–248. Springer.

Basin, D., Debois, S., and Hildebrandt, T. (2018). On purpose and by necessity: compliance under the gdpr. In *International Conference on Financial Cryptography and Data Security*, pages 20–37. Springer.

Bates, A., Tian, D. J., Butler, K. R., and Moyer, T. (2015). Trustworthy whole-system provenance for the linux kernel. In *24th {USENIX} Security Symposium ({USENIX} Security 15)*, pages 319–334.

Bier, C. (2013). How usage control and provenance tracking get together-a data protection perspective. In *2013 IEEE Security and Privacy Workshops*, pages 13–17. IEEE.

Bonatti, P., Kirrane, S., Polleres, A., and Wenning, R. (2017). Transparent personal data processing: The road ahead. In *International Conference on Computer Safety, Reliability, and Security*, pages 337–349. Springer.

Council of European Union (2016). Council regulation (EU) no 2016/679. `https://eur-lex.europa.eu/eli/reg/2016/679/oj`.

Freire, J., Koop, D., Santos, E., and Silva, C. T. (2008). Provenance for computational tasks: A survey. *Computing in Science & Engineering*, 10(3):11–21.

Garijo, D. and Gil, Y. (2013). P-Plan: The P-Plan ontology. W3C recommendation, W3C. https://www.opmw.org/model/p-plan17092013/.

GDPR.EU (2019). 2019 GDPR Small Business Survey: Insights from European small business leaders one year into the General Data Protection Regula-

tion. https://gdpr.eu/wp-content/uploads/2019/05/2019-GDPR.EU-Small-Business-Survey.pdf.

Gjermundrød, H., Dionysiou, I., and Costa, K. (2016). privacytracker: a privacy-by-design gdpr-compliant framework with verifiable data traceability controls. In *International Conference on Web Engineering*, pages 3–15. Springer.

Kuner, C. (2012). The european commission's proposed data protection regulation: A copernican revolution in european data protection law. *Bloomberg BNA Privacy and Security Law Report (2012) February*, 6(2012):1–15.

Martin, A. P., Lyle, J., and Namiluko, C. (2012). Provenance as a security control. In *TaPP*.

Moreau, L. and Missier, P. (2013). PROV-dm: The PROV data model. W3C recommendation, W3C. http://www.w3.org/TR/2013/REC-prov-dm-20130430/.

Ozsoyoglu, G. and Snodgrass, R. T. (1995). Temporal and real-time databases: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 7(4):513–532.

Pandit, H. J. and Lewis, D. (2017). Modelling provenance for gdpr compliance using linked open data vocabularies. In *PrivOn@ ISWC*.

Pandit, H. J., O'Sullivan, D., and Lewis, D. (2019). Test-driven approach towards gdpr compliance. In Acosta, M., Cudré-Mauroux, P., Maleshkova, M., Pellegrini, T., Sack, H., and Sure-Vetter, Y., editors, *Semantic Systems. The Power of AI and Knowledge Graphs*, pages 19–33, Cham. Springer International Publishing.

Pasquier, T. F.-M., Singh, J., Eyers, D., and Bacon, J. (2015). Camflow: Managed data-sharing for cloud services. *IEEE Transactions on Cloud Computing*, 5(3):472–484.

Pohly, D. J., McLaughlin, S., McDaniel, P., and Butler, K. (2012). Hi-fi: collecting high-fidelity whole-system provenance. In *Proceedings of the 28th Annual Computer Security Applications Conference on*, pages 259–268.

Shastri, S., Banakar, V., Wasserman, M., Kumar, A., and Chidambaram, V. (2019). Understanding and benchmarking the impact of gdpr on database systems. *arXiv preprint arXiv:1910.00728*.

Tankard, C. (2016). What the gdpr means for businesses. *Network Security*, 2016(6):5–8.

Ujcich, B. E., Bates, A., and Sanders, W. H. (2018). A provenance model for the european union general data protection regulation. In *International Provenance and Annotation Workshop*, pages 45–57. Springer.

Wang, L., Near, J. P., Somani, N., Gao, P., Low, A., Dao, D., and Song, D. (2019). Data capsule: A new paradigm for automatic compliance with data privacy regulations. In *Heterogeneous Data Management, Polystores, and Analytics for Healthcare*, pages 3–23. Springer.