UNIVERSIDAD POLITÉCNICA DE MADRID
Escuela Técnica Superior de Ingenieros Informáticos

Semantic Representation of Privacy Terms and Policy-based Algorithms for Decentralised Data Environments

DOCTORAL THESIS

Submitted for the degree of Doctor by:

Beatriz Gonçalves Crisóstomo Esteves
Master Degree in Biomedical Engineering

Madrid, 2024

UNIVERSIDAD POLITÉCNICA DE MADRID
Escuela Técnica Superior de Ingenieros Informáticos

Doctoral Degree in Artificial Intelligence

Semantic Representation of Privacy Terms and Policy-based Algorithms for Decentralised Data Environments

DOCTORAL THESIS

Submitted for the degree of Doctor by:

Beatriz Gonçalves Crisóstomo Esteves
Master Degree in Biomedical Engineering

Under the supervision of:
Dr. Víctor Rodríguez Doncel
Dr. David Lewis

Madrid, 2024

Title: Semantic Representation of Privacy Terms and Policy-based Algorithms for Decentralised Data Environments

Author: Beatriz Gonçalves Crisóstomo Esteves

Doctoral Programme: Artificial Intelligence

Thesis Supervision:

Dr. Víctor Rodríguez Doncel, Associate Professor, Universidad Politécnica de Madrid(Supervisor)

Dr. David Lewis, Associate Professor, Trinity College Dublin

External Reviewers:

Thesis Defense Committee:

Thesis Defense Date:

This Thesis has been partially supported by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 813497 (PROTECT).

Acknowledgement

This thesis has been accomplished with the support, assistance and encouragement of many people to whom I would like to express my deepest gratitude.

First of all, I would like to express my gratitude to the Ontology Engineering Group for supporting the work developed in this thesis, in particular to my supervisor Professor Víctor Rodríguez Doncel. Thank you for providing me with the incredible opportunity to explore and develop my own ideas in distinct fields of knowledge and for supporting me throughout this remarkable experience. I would also like to express my gratitude to Professor Elena Montiel Ponsoda for the insightful discussions and the opportunity to work in INESData. Furthermore, I would like to thank my co-supervisor, Professor Dave Lewis, for his enthusiasm, interest, and comprehensive support. Last, but certainly not least, my sincere gratitude goes to Professor Harshvardhan Pandit for being an incredible mentor. Thank you Harsh, for your assistance, discussions, collaborations, technical and legal expertise, as well as just chatting over many offline and online meetings.

Secondly, I would also like to acknowledge the funding received from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 813497 (PROTECT), European Union's Horizon 2020 research and innovation programme under grant agreement No. 101036418 (AURORA), and the NextGenerationEU-funded INESData project. I would like to thank all the fellows and professors involved in these projects. In particular, I am especially grateful to my fellow PROTECT Work Package 1 colleagues, Andrés Chomczyk Penedo, Blessing Mutiro, and Haleh Asgarinia, for the insightful discussions that contributed to the interdisciplinarity tone of this thesis. It was, and still is, a pleasure to work with you.

I would also like to thank the people that made my research stays memorable and kickstarted collaborations that I hope will last for many years. To Dr. Kieran Fraser, my research stay at Empushy was a great opportunity to apply my legal knowledge in a startup environment. To Pat McBennett, our continuous discussions and interactions were immensely helpful in applying this thesis outcomes to relevant industrial settings. To Dave and Harsh, my research stay at Trinity College Dublin was pivotal in the development of PLASMA and DGAterms. To Professor Ruben Verborgh, while short, my research stay at UGent was incredibly insightful. The work being developed in Flanders, centred on Solid, and in particular by the KNoWS group was and is an inspiration that drove the final touches of this thesis. I would also like to extend a final thanks to Ruben for being incredibly kind and providing actionable feedback that immensely contributed to the final text that can be read in this thesis.

To the DPVCG and ODRL CG people, your incentive and feedback were immensely helpful for the validation of this thesis.

À Carina, obrigada pelo suporte, companheirismo e pelas muitas noites passadas a escrever papers e a fazer-me figuras. Nada disto teria sido possível sem ti, obrigada por nunca me teres deixado desistir. *My personal assistant, forever.* Por fim, não posso deixar de agradecer à minha família e amigos: mãe, pai, Nuno, avós, padrinho e madrinha, Gui e Afonso, e todos os outros, muito obrigada! Obrigada a todos pelo apoio, pelo interesse e pela preocupação.

Abstract

With the widespread of technologies in every aspect of our day-to-day life, the amount of data available worldwide is growing rapidly and, consequently, the legal and ethical implications of its exploration have been under debate for quite a few years. When the General Data Protection Regulation (GDPR) came into full effect on the 25th of May 2018, companies had to deal with the impact of this new legislation on their processing of personal data and users were overloaded with the amount of complex technical information on their renewed rights over that processing. The main goal of this thesis is to find ways to help users of Web services deal with this overload, offering services that match their preferences and respect their rights, aiding them in taking control over the publication and sharing of their personal data.

In this context, the use and extension of data protection vocabularies and machine-readable policy languages are suitable for the representation of individual privacy preferences and requirements, fine-grained policies for the processing of personal data and other machine-readable information related to GDPR rights and obligations, including the logging of processing activities for future auditing and the exercising of user's personal data-related rights. Furthermore, these specifications can also be used to establish a policy matching mechanism where fine-grained GDPR-aligned access control policies are used to manage and determine access to decentralised personal datastores, such as Solid Pods. Solid is a decentralised data environment that detaches the storage of data from the processing of said data performed by data-driven applications. Such an architecture allows Web users to have better control over the movement of their personal data and regain trust in the services using it as the users are the ones specifying who can access their data. The policy matching algorithm and the developed vocabularies are also used to deal with the requirements of sharing health data and to manage the requirements of the newly enforced Data Governance Act to showcase the representational capabilities of the developed technologies to cover specific use cases and to be expanded to deal with new demands, in particular, related to the expression of data reuse policies and consent terms.

The contributions proposed in this Thesis confirm the hypothesis that Semantic Web technologies can be used to successfully express data protection-related information, including the definition of data subjects' privacy preferences as access control policies related to their personal data. Furthermore, said technologies can be used to increase the transparency and accountability of decentralised data environments, in particular when it comes to the involved entities and infrastructure, including their access control mechanisms.

Resumen

Con la expansión de las tecnologías en todos los aspectos de nuestra vida cotidiana, la cantidad de datos disponibles en todo el mundo está creciendo rápidamente y, en consecuencia, las implicaciones legales y éticas de su exploración han sido objeto de debate durante bastantes años. Cuando el Reglamento General de Protección de Datos (RGPD) entró en pleno vigor el 25 de mayo de 2018, las empresas tuvieron que lidiar con el impacto de esta nueva legislación en su procesamiento de datos personales y los usuarios se vieron sobrecargados con la cantidad de información técnica compleja relacionada con sus derechos renovados sobre ese tratamiento. El objetivo principal de esta tesis es encontrar formas de ayudar a los usuarios de servicios Web a lidiar con esta sobrecarga, ofreciéndoles servicios que se ajusten a sus preferencias y respeten sus derechos, ayudándoles a tomar control sobre la publicación y el intercambio de sus datos personales.

En este contexto, el uso y la ampliación de vocabularios de protección de datos y lenguajes de políticas son adecuados para la representación de preferencias y requisitos de privacidad individuales, políticas detalladas para el procesamiento de datos personales y otra información legible por máquinas relacionada con los derechos y obligaciones del RGPD, incluido el registro de las actividades de procesamiento para futuras auditorías y el ejercicio de los derechos del usuario relacionados con los datos personales. Además, estas especificaciones también se pueden utilizar para establecer un mecanismo de coincidencia de políticas en el que se utilicen políticas de control de acceso detalladas y alineadas con el RGPD para gestionar y determinar el acceso a almacenes de datos personales descentralizados, como Solid Pods. Solid es un ambiente de datos descentralizado que separa el almacenamiento de datos del procesamiento de dichos datos realizado por aplicaciones. Esta arquitectura permite a los usuarios de la Web tener un mejor control sobre el movimiento de sus datos personales y recuperar la confianza en los servicios que los utilizan, ya que son los usuarios quienes especifican quién puede acceder a sus datos. El algoritmo de coincidencia de políticas y los vocabularios desarrollados también se utilizan para abordar los requisitos de compartir datos de salud y para gestionar los requisitos de la Ley de Gobernanza de Datos recientemente aplicada para mostrar las capacidades de representación de las tecnologías desarrolladas para cubrir casos de uso específicos y ampliarse para hacer frente a nuevas demandas, en particular, relacionadas con la expresión de políticas de reutilización de datos y términos de consentimiento.

Las contribuciones propuestas en esta Tesis confirman la hipótesis de que las tecnologías de la Web Semántica pueden usarse para expresar con éxito información relacionada con la protección de datos, incluida la definición de las preferencias de privacidad de los interesados como políticas de control de acceso relacionadas con sus datos personales. Además, dichas tecnologías se pueden utilizar para aumentar la transparencia y la rendición de cuentas de los ambientes de datos descentralizados, en particular cuando se trata de las entidades y la infraestructura involucradas, incluidos sus mecanismos de control de acceso.

Table of Contents

List of Figures

List of Tables

List of Listings

List of Algorithms

Abbreviations and acronyms

ACL	Access Control List
ACP	Access Control Policy
ACR	Access Control Resource
API	Application Programming Interface
APN	Annotation of Push-Notifications ontology
ASP	Answer Set Programming
CG	Community Group
CQ	Competency Question
CSS	Community Solid Server
DATS	Data Tags Suite
DGA	Data Governance Act
DID	Decentralized Identifier
DKG	Distributed Knowledge Graphs
DNT	Do Not Track
DPIA	Data Protection Impact Assessment
DPO	Data Protection Officer
DPV	Data Privacy Vocabulary
DPVCG	Data Protection Vocabularies and Controls Community Group
DUL	Data Use Limitation
DUO	Data Use Ontology
DUODRL	ODRL for DUO vocabulary
DUOS	Data Use Oversight System
EC	European Commission
ECHR	European Convention on Human Rights
EDIB	European Data Innovation Board

EDPB European Data Protection Board

EDPS European Data Protection Supervisor

EHDS European Health Data Space

eIDAS electronic IDentification, Authentication and trust Services

ENISA European Union Agency for Cybersecurity

ESS Enterprise Solid Server

EU European Union

FAIR Findable, Accessible, Interoperable and Reusable

FIPPs Fair Information Practice Principles

FOOPS! OntOlogy Pitfall Scanner for FAIR

GA4GH Global Alliance for Genomics and Health

GDPR General Data Protection Regulation

IEC International Electrotechnical Commission

IoT Internet of Things

ISO International Organization for Standardization

LDN Linked Data Notification

LOT Linked Open Terms

LOV Linked Open Vocabularies

OAC ODRL profile for Access Control

OBI Ontology for Biomedical Investigations

OOPS! OntOlogy Pitfall Scanner

ORSD Ontology Requirement Specification Document

P3P Platform for Privacy Preferences

PAV Provenance, Authoring and Versioning

PET Privacy Enhancing Technology

PII Personally Identifiable Information

PIMS Personal Information Management System

PLASMA Policy LAnguage for Solid's Metadata-based Access control

PoC Proof of Concept

OBO Open Biological and Biomedical Ontology

SAI Solid Application Interoperability

SAR Subject Access Request

SoDA Solid for Data Altruism

SOPE Solid ODRL access control Policies Editor

TOM Technical and Organisational Measure

UDHR Universal Declaration of Human Rights

UML Unified Modeling Language

UI User Interface

VC Verifiable Credential

W3C World Wide Web Consortium

WAC Web Access Control

WP 29 Article 29 Data Protection Working Party

Namespaces

PREFIX access-right: <http://purl.org/coar/access_right/>

PREFIX acl: <http://www.w3.org/ns/auth/acl#>

PREFIX acp: <http://www.w3.org/ns/solid/acp#>

PREFIX cc: <http://creativecommons.org/ns#>

PREFIX dcat: <http://www.w3.org/ns/dcat#>

PREFIX dcterms: <http://purl.org/dc/terms/>

PREFIX dgaterms: <https://w3id.org/dgaterms#>

PREFIX dpo: <http://www.uni.lu/dataprotection#>

PREFIX dpv: <https://w3id.org/dpv#>

PREFIX duo: <http://purl.obolibrary.org/obo/duo.owl#>

PREFIX duodrl: <https://w3id.org/duodrl#>

PREFIX eu-gdpr: <https://w3id.org/dpv/legal/eu/gdpr#>

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

PREFIX foaf: <http://xmlns.com/foaf/0.1/>

PREFIX gc: <https://w3id.org/GConsent#>

PREFIX gdprov: <https://w3id.org/GDPRov#>

PREFIX interop: <http://www.w3.org/ns/solid/interop#>

PREFIX justif: <https://w3id.org/people/besteves/justifications#>

PREFIX ldp: <http://www.w3.org/ns/ldp#>

PREFIX legal-eu: <https://w3id.org/dpv/legal/eu#>

PREFIX lkif: <http://www.estrellaproject.org/lkif-core/lkif-core.owl#>

PREFIX oac: <https://w3id.org/oac#>

PREFIX obo: <http://purl.obolibrary.org/obo/>

PREFIX odrl: <http://www.w3.org/ns/odrl/2/>

PREFIX p3p: <http://www.w3.org/2002/01/P3Pv1#>

PREFIX pav: <http://purl.org/pav/>

PREFIX pd: <https://w3id.org/dpv/pd#>

PREFIX pim: <http://www.w3.org/ns/pim/space#>

PREFIX plasma: <https://w3id.org/plasma#>

PREFIX prov: <http://www.w3.org/ns/prov#>

PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

PREFIX schema: <https://schema.org/>

PREFIX sh: <http://www.w3.org/ns/shacl#>

PREFIX skos: <http://www.w3.org/2004/02/skos/core#>

PREFIX solid: <http://www.w3.org/ns/solid/terms#>

PREFIX tech: <https://w3id.org/dpv/tech#>

PREFIX ti: <http://www.ontologydesignpatterns.org/cp/owl/timeinterval.owl#>

PREFIX time: <http://www.w3.org/2006/time#>

PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

Part I

INTRODUCTION

Chapter 1

Introduction

As Westin [1967a] predicted in his *"Legal Safeguards to Insure Privacy in a Computer Society"*, the rapid development of data surveillance technology overpowered the individual right to privacy in favour of business profit. His postulates on privacy, written before the inventions of the internet and the Web, influenced the privacy regulations enacted in the following decades and its impact can still be seen in the European Union's (EU) General Data Protection Regulation (GDPR), through its *actionable* data subject rights, which intend to give individuals the freedom to control their own personal information [Westin, 1967b]. Throughout his career, from 1990 to 2003, Westin also conducted a series of surveys related to consumer attitudes towards privacy, where a majority of consumers reported having *"lost all control over how personal information about them is circulated and used by companies"* [Kumaraguru and Cranor, 2005]. Additionally, these surveys highlighted the variety of individual concerns about privacy – the 1996's study showed that 25% of the public are fundamentalists (people who are highly concerned with their privacy), 59% pragmatists (people who are concerned with their privacy and want to protect themselves from the abuse or misuse of their personal information by companies or government agencies) and 16% unconcerned (people who have no real concerns about privacy) and a similar study in 2003 [Taylor, 2003] showed an increase in the percentage of pragmatists, 64%, and a decrease in the unconcerned, 10% – a powerful reminder that, while there are more people with real concerns about the misuse of their data, different people want to have a different level of control over their privacy settings.

Following Westin's postulates, Convention 108 [Council of Europe, 1981] was created as the first and only international data protection instrument, and over the years, enhancements have been made to address automated data processing, and trans-jurisdictional data flows. Furthermore, in Europe, the first data protection law, the Data Protection Directive, was introduced in 1995 and the Charter of Fundamental Rights acknowledged the *"right to the protection of personal data"* in 2000. This last one was revised and expanded in 2016 to deal with the intensive usage of digital personal data, in the form of the GDPR [2016b]. Building on this, in 2020, and after widespread adoption of GDPR-like data protection regulations around the world [Bradford, 2019], the European Commission launched its *strategy for data* [European Commission, 2020], with the goal of allowing the cross-sector flow of data within the EU, while ensuring that the *"European rules and values, in particular personal data protection, consumer protection legislation and competition law, are fully respected"*. Since then, we have seen the launch of a series of proposals for regulation by the

European Commission and Parliament, some of them already approved and being enforced, that build on the GDPR when it comes to the processing of personal data, with the main purpose of promoting the EU's data economy while better placing EU's citizens in control of what happens to their data. In parallel, technological development with similar goals of empowering data subjects through increased retention and control over their data has grown, however these two strands have not been well integrated to date.

1.1 Thesis Overview

The content of this Thesis is based on the research work, and corresponding published articles, developed from January 2020 to December 2023. ChatGPT was not used for the development of this Thesis.

Part I: Introduction

This Chapter presents the motivation of this Thesis and resulting publications, a set of definitions, as well as the projects and research stays accomplished during the Thesis.

Chapter 2 provides a state of the art on (i) decentralising the access to personal data with Solid, (ii) representing personal data processing information in a machine-readable format, and (iii) using policy languages to specify access control conditions.

Chapter 3 describes the objectives, hypotheses, assumptions, restrictions, research questions, contributions, and methodology followed throughout this Thesis.

Part II: GDPR-Aligned Vocabularies for Personal Datastores

Chapter 4 describes the developed vocabularies, including (i) an ODRL profile for Access Control (OAC), (ii) a metadata language for Solid (PLASMA), and (iii) rights exercising records using DPV, and includes the ontologies quality evaluation, comprising the detection of common pitfalls, validation of competency questions with SPARQL queries, alignment with FAIR principles and with the International Organization for Standardization/International Electrotechnical Commission (ISO/IEC) 27560 standard.

Chapter 5 includes a legal and ethical discussion, including collaborations with the law and ethics experts in PROTECT, and other EU-funded projects, as well as with the law experts in the W3C DPVCG.

Part III: Algorithms & Use Cases

Chapter 6 describes the policy matching algorithm and presents a proof of concept implementation that uses and extends the developed vocabularies to deal with the specific requirements of health data sharing. Proof of concept implementations to specify privacy preferences and to exercise the right of access are also presented.

Chapter 7 presents a use case validation of this work by applying the developed resources as a building block for the creation of policies for the new Data Governance Act (DGA).

Part IV: Conclusions

Chapter 8 describes the conclusions of this Thesis and presents future work.

1.2 Motivation

GDPR [2016b] came into full effect on the 25th of May 2018 with the main objective of providing the European Union's natural persons with the right to protection of their personal data, especially in relation to its fair, transparent, and lawful processing and sharing, including a series of rights regarding portability and erasure of data or objection to processing [Ausloos et al., 2019]. This Regulation revolves around the relationship between *'data subjects'* – the natural personal to which the personal data refers/identifies – and *'data controllers'* – the legal entities processing said personal data. As previously mentioned, a large part of controllers' compliance obligations are related to the data subject's rights defined in Chapter III of the GDPR. Information related to these rights should be provided to data subjects in a concise, transparent, comprehensible, and easily accessible manner, as well as in clear and plain language. In particular, the so-called *'Right to be informed'*, described in Articles 13 and 14, establishes the information that should be provided to the data subject at the time when the data is first collected, e.g., information on the identity of the controllers, purposes for processing, information on data transfers or existence of automated decision-making. Moreover, data subjects should also be provided with information regarding their other rights: the right of access to the personal data being processed; the right of rectification of inaccurate personal data; the right to be forgotten, i.e., the data controller has to erase the personal data requested by the subject; the right to restrict the processing of personal data; the right to be notified about the rectification, erasure, or restriction of processing; the right to data portability; the right to object to any processing, including profiling; and the right to not be subjected to automated decision-making, including profiling.

Companies usually deal with these information requirements by providing a description of their personal data-handling services in their privacy policies [Linden et al., 2020], however, these are usually difficult to comprehend due to their complexity, lack of readability, and usage of legal terms [McDonald and Cranor, 2008, Pasquale, 2015, Fabian et al., 2017, Lovato et al., 2023], which lead Web users to ignore them for the sake of having access to the Web services they want to use [Gindin, 2009, Rudolph et al., 2018, Obar and Oeldorf-Hirsch, 2020]. As such, machine-readable policy languages seem perfectly fitted to convey these information requirements in a more transparent manner – they have been on the Web scene since the 1990s with the primary goal of establishing the conditions to access Web resources [Zhao et al., 2016, Pellegrini et al., 2018b, Leicht and Heisel, 2019], they can offer different ways of examining privacy policies through different user interfaces [Angulo et al., 2012, Gerl et al., 2020], independently from the data controllers writing said policies, and they already can encode some privacy terms [Cranor et al., 2002b, Iannella and Villata, 2018], such as the purpose for processing or recipients. On the other hand, they are not enough to invoke specific legal terms related to the information that must be shared by data controllers, such as the legal basis for processing or the existing data subject rights. In this context, a new wave of Semantic Web privacy and data protection vocabularies and ontologies has appeared, which can be used to represent this information, no doubt due to the proliferation of the GDPR and other data privacy-related laws [Pandit, 2020, Esteves and Rodríguez-Doncel, 2022a]. Thus,

such policy languages and vocabularies can be proven useful to assist data controllers in achieving GDPR alignment for their Web services and to help data subjects in the management of their rights, whether related to the transparent information requirements or their other GDPR-based rights.

More than that, the Semantic Web domain itself is of extreme importance for the representation of these privacy terms as it drives the development of open standards and specifications with interoperability and extensibility in mind. This effort was and is being led by the World Wide Web Consortium (W3C), openly and collaboratively, with the cooperation of academia and industry. In this context, two main lines of work are pursued: (i) the development of common formats for data interoperability to ensure seamless integration of data from distinct sources; and (ii) the promotion of a structured language with the ability to document how data relates to real-world objects. Semantic Web technologies can therefore be applied to a wide range of application fields: data integration, improved search, content management and discovery, domain modelling, or semantic annotation [2012]. The term *'Semantic Web'* was first formulated by Sir Tim Berners-Lee, Web inventor and founder of the W3C, with the goal of having a *'Web of Data'*, an extension of the Web of Documents so that data can be shared and reused in a granular manner across applications, companies and the Web community in general [Berners-Lee et al., 2001]. Solid emerged then as a natural solution to deliver this promise as a Web standards-based decentralised storage environment for data with an integrated, granular access control mechanism [Sambra et al., 2016, Verborgh, 2022]. Thus, such a system allows its users to choose who has access to their data and what applications to use, fulfilling GDPR's requirements of improving data portability and control for data subjects. This implies a significant shift in the *status quo* where companies gather and process data on a massive scale, constrained only by what they state in their privacy policies [Nixdorf, 2019, Robertson, 2020], where they can control what and how such information is stated. As such, for Solid to be regarded as a tool that is fully aligned with the GDPR, the information requirements and other previously mentioned rights need to be considered and integrated into the Solid platform, in the form of user policies, notices, data access agreements, registers of users and applications, and logging of Solid-related activities.

1.3 Definitions

In this Section, a series of terms, which are used throughout this Thesis, is introduced.

1.3.1 Privacy terms

The expression *'privacy terms'* is usually associated with the notice that personal data handling entities use to disclose how they collect, use, store, secure, and share personal data (1.3.5). In this Thesis, it has a broader meaning – the expression *'privacy terms'* will be used to refer to the information that needs to be modelled to represent concepts related to the preferences and rights of data subjects and to the policies (1.3.2) and obligations of data controllers regarding privacy and personal data protection [Esteves, 2021]. Examples of privacy terms are the *purpose* for processing personal data, the *processing operation* itself, the *legal basis* used by the data controller to justify the processing, or the *right to data portability*.

1.3.2 Policies

Policies are documents that describe conditions for access and usage of content. Such policies can be expressed as permissions, prohibitions, or obligations and include constraints, i.e., conditions that refine the policy rules [Iannella and Villata, 2018]. In this Thesis, the focus will be on the design of policies to determine access to personal data assets stored in decentralised data systems. Examples of these policies are *user preferences and requirements*, *applications privacy policies* and *data requests* or *data access agreements* regarding data handling practices over personal data stored or shared through personal datastores (1.3.6).

1.3.3 Access control

The term *'access control'* refers to the model used to guide the process of access to resources. To do so, access control rules can be defined through a policy language (1.3.2) with specific syntax and semantics, which are the base of a policy enforcement mechanism [Kirrane et al., 2017]. Moreover, both authentication and authorisation mechanisms – processes related to identity verification and rule-based access control enforcement, respectively – are involved in the process of granting or denying access to resources. In simple terms, an access control policy involves three aspects: (i) the entity requesting/being requested access; (ii) the requested resource(s); and (iii) the access rules, i.e., permissions and prohibitions on particular access modes.

1.3.4 Legislation on data protection

Before data protection was considered a fundamental right, the right to privacy emerged in 1948 through the Universal Declaration of Human Rights (UDHR) under Article 12, which stated that *"No one shall be subjected to arbitrary interference with his privacy, family, home or correspondence [...]"* [United Nations General Assembly, 1948], and through the European Convention on Human Rights (ECHR) under Article 8 as *"Everyone has the right to respect for his private and family life, his home and his correspondence"* [Council of Europe, 1950].

The first and only legally binding international instrument in the field of data protection, Convention 108 [Council of Europe, 1981], was created by the Council of Europe in 1981 and has been improved throughout the years to deal with automatic processing, supervisory authorities, and transborder data flows. Following in these footsteps, the Charter of Fundamental Rights of the EU [2000], under Article 8, acknowledged that *"Everyone has the right to the protection of personal data concerning him or her"* and that *"Such data must be processed fairly for specified purposes and on the basis of the consent of the person concerned or some other legitimate basis laid down by law"*, in addition to the right to privacy laid down in Article 7. It should also be noted that in Spain, the right to privacy, or *'derecho a la intimidad'*, is also a fundamental right since 1978, as defined in Article 18 of the Spanish Constitution [1978].

The first EU law on data protection, the Data Protection Directive, was launched in 1995. Further legislation was enforced in 2018, to adapt the EU law on data protection to the intensive usage of digital personal data (1.3.5) by new technological developments, in the form of the GDPR [2016b]. In the same package, the EU launched a similar piece of legislation for the processing of personal data by state authorities for law enforcement purposes, i.e., preventing, investigating, detecting,

and prosecuting criminal offences or executing criminal penalties, the Directive 2016/680 [2016a].

This Thesis focuses on the requirements brought by the GDPR (1.3.4) and also resorts to the opinions and guidelines adopted by the Article 29 Working Party, the European Data Protection Board (EDPB) and the European Data Protection Supervisor (EDPS), as well as case law and other legal literature [European Union Agency for Fundamental Rights and Council of Europe, 2018].

GDPR

GDPR [2016b] is the landmark legislation on data protection in Europe and its effects have been felt throughout the globe, with similar pieces of regulation being discussed and adopted in American, African, and Asian countries [Bradford, 2019]. It established the following principles related to the processing of personal data:

(i) it should be processed in a legal, fair, and transparent manner (Article 5.1(a) on *'lawfulness, fairness and transparency'*).

(ii) the processing should be limited to the purpose specified by the data subject (Article 5.1(b) on *'purpose limitation'*).

(iii) it should include only the minimum data relevant to the purpose for which it is being processed (Article 5.1(c) on *'data minimisation'*).

(iv) the data should be accurate and possible to be corrected when necessary (Article 5.1(d) on *'accuracy'*).

(v) it should be kept only as long as it is necessary (Article 5.1(e) on *'storage limitation'*).

(vi) adequate technical and organisational measures for the security of the personal data should be ensured (Article 5.1(f) on *'integrity and confidentiality'*).

(vii) data controllers should have accountability mechanisms to demonstrate compliance with these principles (Article 5.2 on *'accountability'*).

Moreover, while analysing Chapters III and IV (*'Rights of the data subject'*[1] and *'Controller and processor'*[2], respectively) of the GDPR, a set of information flows between data-related entities, i.e., data subjects, data controllers, processors and data protection officers, recipients, or supervisory authorities, can be identified. In this context, an information flow refers to the information that has to be shared between entities so that a right or obligation can be exercised or fulfilled and GDPR's principles of lawfulness, fairness, and transparency can be respected. As such, the following rights are provided to data subjects:

(Arts. 13 and 14) *'right to be informed'* obliges data controllers to inform data subjects about any processing of personal data, whether being from data collected directly from the data subjects or other sources.

(Art. 15) *'right of access'* to the personal data being processed, including a copy of the data as well as information about the purposes for processing, categories of the personal data being

[1] https://gdpr-info.eu/chapter-3/ (accessed on 15 March 2024)
[2] https://gdpr-info.eu/chapter-4/ (accessed on 15 March 2024)

processed and their source, any recipients to whom the data may have been shared and corresponding measures to ensure its security, storage and retention conditions and the existence of other data subject's rights.

(Art. 16) *'right to rectification'* of inaccurate or incomplete personal data.

(Art. 17) *'right to be forgotten'* by the data controllers when the personal data is no longer needed for the purposes for which it was collected.

(Art. 18) *'right to restriction of processing'* of personal data when its accuracy is being contested, the processing is unlawful, when the data subject needs it for any legal claims or objects to the processing.

(Art. 19) *'right to be notified'* about the rectification, erasure or restriction of processing.

(Art. 20) *'right to data portability'*, including the right to request that said data be transferred directly from one controller to another.

(Art. 21) *'right to object'* to any processing, including profiling.

(Art. 22) *'right to not be subjected to automated decision-making'*, including profiling.

For instance, if a data subject wishes to exercise its *'right to be forgotten'*, or *'right to erasure'*, apart from raising such a request, there is the need to represent information related to the grounds on which the request is based, and the data controller needs to forward this request to other controllers which are processing the same personal data. Figure 1.1 shows a diagram of such information flows. While the focus of this Thesis relies on the fulfilment of data subjects' rights, it is important to notice as well that the GDPR also contains other provisions, including obligations on data controllers and processors to maintain records of their personal data processing activities, requirements on transfers of personal data to third countries or international organisations, or on the activities of data protection supervisory authorities.

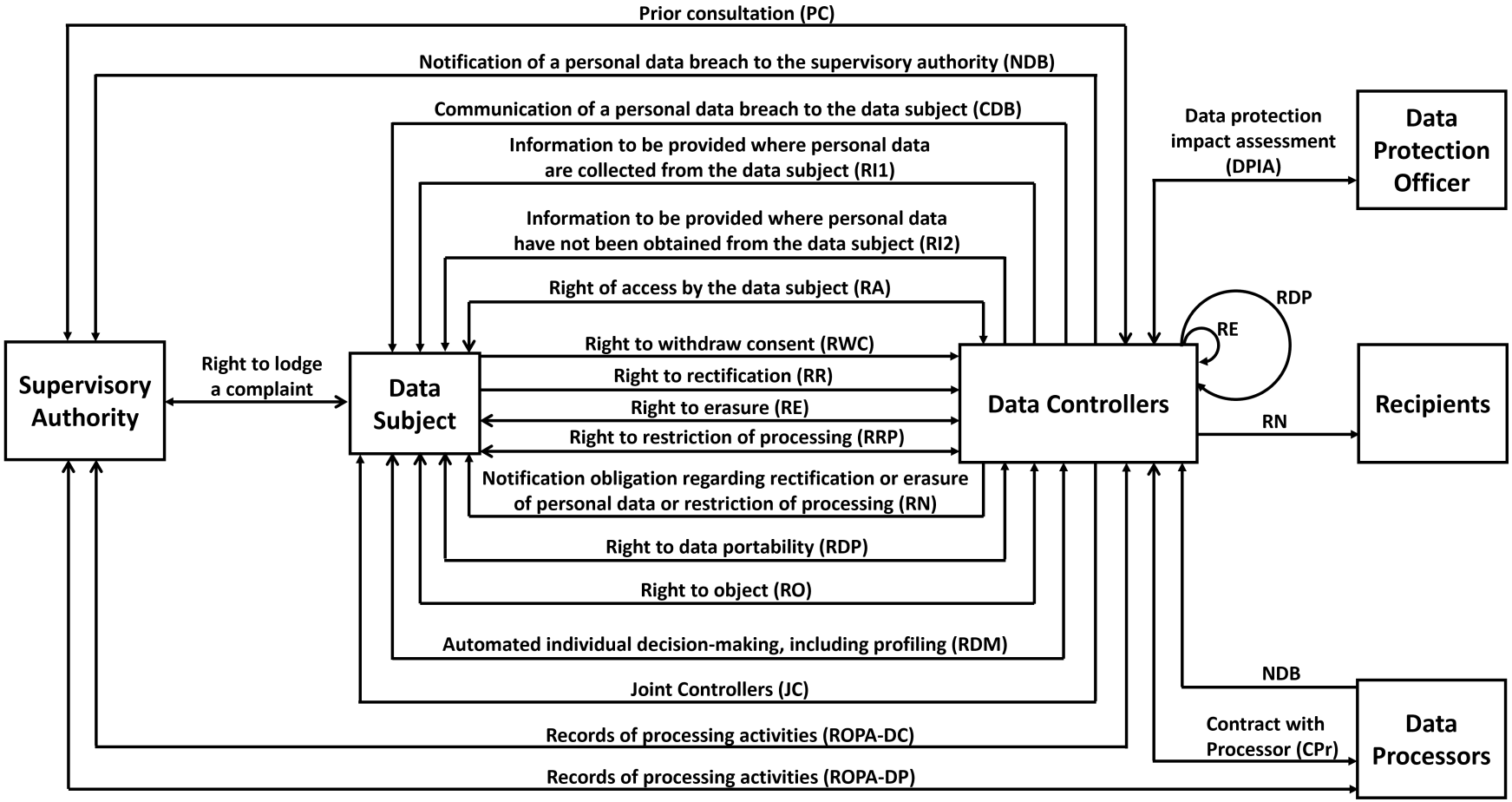


Figure 1.1: GDPR's rights and obligations as information flows. The bidirectional arrows represent a right or obligation in which a request for information and respective response is expected (the open arrowhead, ➜, represents the entity waiting for the response, and the closed arrowhead, ➡, the entity being requested). In contrast, the unidirectional arrows represent only a request or notification and no reply is expected (with the closed arrowhead, ➡, representing the entity being requested or notified), adapted from Esteves and Rodríguez-Doncel [2022a].

1.3.5 Personal data

GDPR's Article 4.1 [2016b] defines 'personal data' as *"any information relating to an identified or identifiable natural person"* such as *"a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person"*. This Thesis relies on GDPR's definition of personal data and special categories of personal data. In addition, it is also acknowledged that there are categories of data that are sensitive, even though they are not considered 'special' under GDPR's Article 9.1, which might require additional consideration and/or protection, e.g., location data has the potential to reveal religious beliefs, sexual orientation or political opinions[3].

Table 1.1, derived from the analysis of Rumbold and Pierscionek [2018], illustrates data sensitivity for particular subcategories of data. For each data type, sensitivity is assessed from 0 to 10, i.e., from low to high sensitivity, and the relative frequency of falling into that particular sensitivity value, on a scale of 1 to 4, is also presented. For instance, data related to objects has a sensitivity of 0 and a frequency of 4 as it can not be used to identify a person, while occupation data would not be frequently classified as being highly sensitive, as demonstrated by the frequency value of 1 to the sensitivity value of 10. In addition, data sensitivity assessment is highly contextual, as shown through the fact that the same data category can have a spectrum of sensitivity, e.g., anonymised data can have a low/high sensitivity if the risk of re-identification is very low/high, respectively. This evaluation of data sensitivity is also well aligned with GDPR's special categories of personal data, as can be seen through the last column of the table, where these special categories are identified and, indeed, the most sensitive are presented in Article 9, with the exception of social class data.

1.3.6 Decentralised data environments

A decentralised environment for data represents a significant paradigm shift in relation to the current status of digital data management. The Web we have today is a centralised Web where data is kept in data silos, controlled only by a handful of Big Tech players, while in a decentralised Web approach *people choose where they store their data* and *exert control over whom gets access to which parts of their data* [Verborgh, 2017].

Figure 1.2 illustrates the distinction between these two paradigms. In a centralised setting, data and applications are coupled and data is kept in *'walled gardens'* controlled by the entities behind centralised platforms such as Facebook, Twitter, or LinkedIn, leaving users without the possibility of reusing it elsewhere [2008a]. By shifting to a decentralised Web, users are able to choose where their data is stored and are in control of their identity, while applications are detached from data, becoming *"views"* over it and fostering innovation and competition through separate markets for data and services [Verborgh, 2022]. Modern decentralised environments include Internet of Things (IoT) ecosystems or personal datastores (1.3.6).

This Thesis focuses on providing users with the tools to better determine access to personal data

[3] DPV models `https://w3id.org/dpv#SensitivePersonalData` as a subtype of `https://w3id.org/dpv#PersonalData` and `https://w3id.org/dpv#SpecialCategoryPersonalData` as a subtype of `https://w3id.org/dpv#SensitivePersonalData`

Table 1.1: Data sensitivity chart derived from Rumbold and Pierscionek [2018]. GDPR's special categories of personal data are identified in the **GDPR** column.

DATA TYPE		DATA SENSITIVITY 10	9	8	7	6	5	4	3	2	1	0	GDPR
Non-personal data	Relating to objects											4	
	Anonymised data related to persons			1	1	1	1	2	3	3	4	3	
Human demographics, behaviour, thoughts & opinions	Opinions				3	3	3	3	3	3	3		
	Purchasing habits			1	2	3	3	3	3	3	3		
	Sex							3	3	3			
	Age				3	3	3	3	3				
	Income			2	3	3	3	3	3				
	Location		1	2	3	3	3	3	3				
	Lifestyle or wellness data		3	3	3	3	3	3	3				
	Occupation	1	2	3	3	3	3	3	3				
	Address			2	3	3	3	3					
	Race			3	3	3	3	3					Art. 9
	Ethnic group			3	3	3	3	3					Art. 9
	Religious or political beliefs			3	3	3	3	3					Art. 9
	Sexual orientation			3	3	3	3	3					Art. 9
	Pregnancy		3	3	3	3	3	3					Art. 9
	Transgender status		3	3	3	3	2						Art. 9
	Social class			3	3	3							
Biometrics	Facial images				3	3	3	3	3	2			Art. 9
	Body images		3	3	3	3	3	3	3	2			Art. 9
	Any traits processed for biometrics		3	3	3								Art. 9
Medical or health data	Diagnoses		3	3	3	3							Art. 9
	Genetic data	3	3	4	3	3							Art. 9
	Highly sensitive diagnoses	4	3										Art. 9

resources stored in decentralised settings, according to EU law on personal data protection. Beyond access control (1.3.3), usage control solutions also need to be developed to provide data subjects with *"control over data usage once access to the data has been granted"* [Akaichi, 2022].

Personal datastores

In its *TechDispatch #3/2020* [European Data Protection Supervisor, 2021], the EDPS envisioned the development of personal data spaces, managed through Personal Information Management Systems (PIMS) as a mechanism to enable personal data sovereignty where *"Individuals, service providers and applications would need to authenticate to access a personal storage centre"* and individuals are able to *"customize what categories of data they want to share and with whom"* while keeping a record of *"who has had access to their digital behaviour"* and enabling data portability and interoperability.

Such decentralised systems (1.3.6) allow data subjects to directly determine who has access to their data, and under which conditions, and can actually play an important role in facilitating the exercise of data subjects' rights, including the rights of access, erasure, and data portability or the right to withdraw consent [Janssen et al., 2020]. In the last few years, different personal datastores initiatives have been gaining prominence and adoption, including the Solid project[4] [Fallatah et al., 2023], which is the adopted use case for the research in this Thesis.

[4] https://solidproject.org/ (accessed on 15 March 2024)

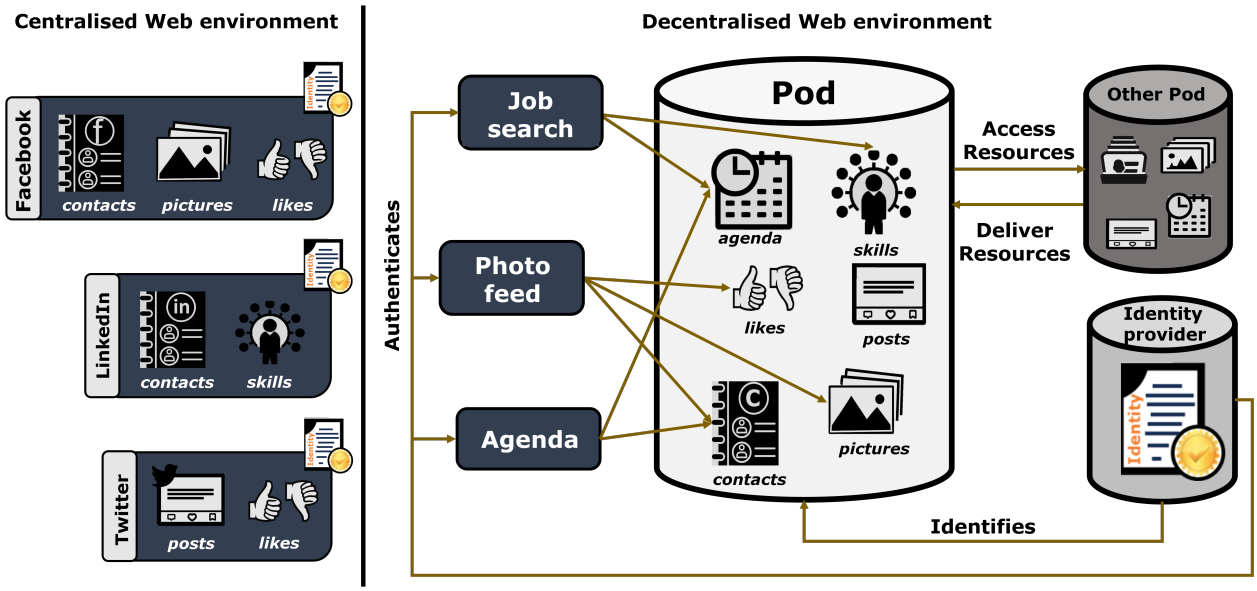


Figure 1.2: Distinction between centralised and decentralised data environments.

Solid is a free, open-source initiative that delivers on the promise of decentralising the storage of data by relying on Web standards and on Semantic Web vocabularies to promote data and services interoperability. To fulfil this vision, the Solid specification relies on authentication and authorisation protocols to provide private, secure, and granular access to data stored in Solid's personal online datastores, the so-called 'Pods' [Mansour et al., 2016].

Moreover, beyond personal datastores, decentralised initiatives at the community level are also being proposed. For instance, data cooperative[5] infrastructures still give their members decision-making control over their data, while allowing them to get paid to share their data in an environment where they have more decision power than what they would have on their own or in other types of data-sharing environments [Mechant et al., 2021]. These community-level stores are also starting to be regulated, e.g., by the DGA [2022g].

1.4 Publications

The following Sections list the works published and presented during the accomplishment of this Thesis. When identified with a †, the authors contributed equally to the publication work.

In addition, Figure 1.3 illustrates the timeline of publications, presentations, and research stays of this Thesis.

1.4.1 Journal contributions

(PJ1) Analysis of Ontologies and Policy Languages to Represent Information Flows in GDPR. (2022) **B. Esteves**, V. Rodríguez-Doncel. *Semantic Web Journal*, pp. 1–35, https://doi.org/10.3233/SW-223009.

[5]Data infrastructure formed through the voluntary and collaborative pooling efforts of individuals.

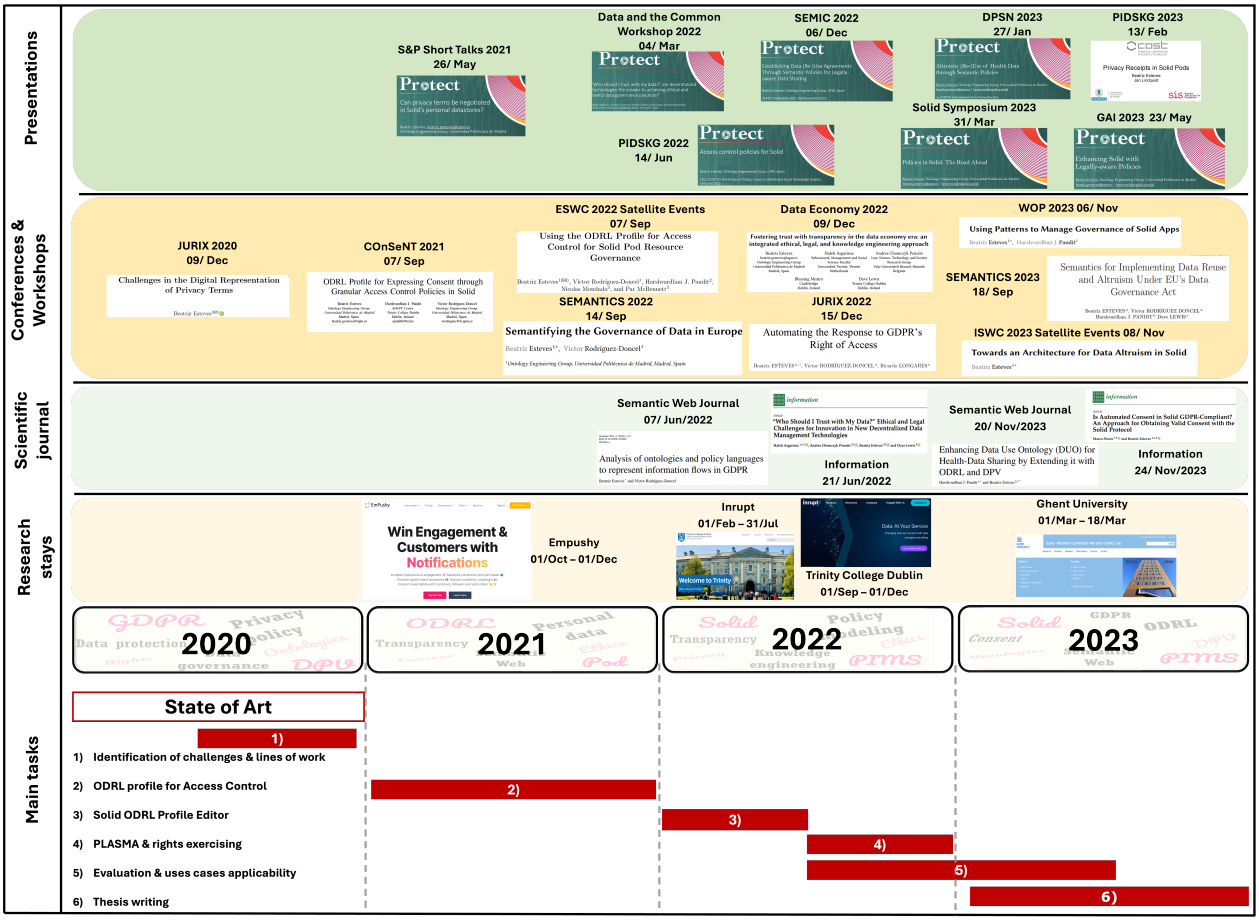


Figure 1.3: Timeline of publications, presentations, and research stays of this Thesis.

(PJ2) "Who Should I Trust with My Data?" Ethical and Legal Challenges for Innovation in New Decentralized Data Management Technologies. (2023) H. Asgarinia†, A. Chomczyk Penedo†, **B. Esteves**†, D. Lewis. *Information 14(7)*, https://doi.org/10.3390/info14070351.

(PJ3) Is Automated Consent in Solid GDPR-Compliant? An Approach for Obtaining Valid Consent with the Solid Protocol. (2023) M. Florea†, **B. Esteves**†. *Information 14(12), Special Issue on Addressing Privacy and Data Protection in New Technological Trends*, https://doi.org/10.3390/info14120631.

(PJ4) Enhancing Data Use Ontology (DUO) for Health-Data Sharing by Extending it with ODRL and DPV. (2023) H. J. Pandit†, **B. Esteves**†. *Semantic Web Journal*, pp. 1–26, https://doi.org/10.3233/SW-243583.

1.4.2 Conference contributions

(PC1) Extracting and Understanding Call-to-actions of Push-Notifications. (2022) **B. Esteves**, K. Fraser, S. Kulkarni, O. Conlan, V. Rodríguez-Doncel. In *Natural Language Processing*

and Information Systems. Edited by P. Rosso, V. Basile, R. Martínez, E. Métais, F. Meziane, Volume 13286, pp. 147–-159. Springer International Publishing. https://doi.org/10.1007/978-3-031-08473-7_14.

(PC2) Now, Later, Never: A Study of Urgency in Mobile Push-Notifications. (2022) **B. Esteves**, K. Fraser, S. Kulkarni, O. Conlan, V. Rodríguez-Doncel. In *Advances in Mobile Computing and Multimedia Intelligence. Edited by P. Delir Haghighi, I. Khalil, G. Kotsis*, pp. 38-–44. Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-20436-4_4.

(PC3) Automating the Response to GDPR's Right of Access. (2022) **B. Esteves**, V. Rodríguez-Doncel, R. Longares. In *Legal Knowledge and Information Systems*, pp. 170—175. IOS Press. https://doi.org/10.3233/FAIA220462.

(PC4) Semantics for Implementing Data Reuse and Altruism Under EU's Data Governance Act. (2023) **B. Esteves**, V. Rodríguez-Doncel, H. J. Pandit, D. Lewis. In *Knowledge Graphs: Semantics, Machine Learning, and Languages. Edited by M. Acosta et al.*, pp. 210–226. IOS Press. https://doi.org/10.3233/SSW230015.

1.4.3 Workshop contributions

(PW1) Challenges in the Digital Representation of Privacy Terms. (2021) **B. Esteves**. In *AI Approaches to the Complexity of Legal Systems XI-XII. Edited by V. Rodríguez-Doncel, M. Palmirani, M. Araszkiewicz, P. Casanovas, U. Pagallo, G. Sartor, Volume 13048*, pp. 313—327. Springer International Publishing. https://doi.org/10.1007/978-3-030-89811-3_22.

(PW2) ODRL Profile for Expressing Consent through Granular Access Control Policies in Solid. (2021) **B. Esteves**, H. J. Pandit, V. Rodríguez-Doncel. In *2021 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, pp. 298-–306. https://doi.org/10.1109/EuroSPW54576.2021.00038.

(PW3) Using the ODRL Profile for Access Control for Solid Pod Resource Governance. (2022) **B. Esteves**, V. Rodríguez-Doncel, H. J. Pandit, N. Mondada, P. McBennett. In *The Semantic Web: ESWC 2022 Satellite Events. Edited by P. Groth, A. Rula, J. Schneider, I. Tiddi, E. Simperl, P. Alexopoulos, R. Hoekstra, M. Alam, A. Dimou, M. Tamper*, pp. 16-–20. Springer International Publishing. https://doi.org/10.1007/978-3-031-11609-4_3.

(PW4) Semantifying the Governance of Data in Europe. (2022) **B. Esteves**, V. Rodríguez-Doncel. In *18th International Conference on Semantic Systems - CEUR Workshop Proceedings, Volume 3235*. https://ceur-ws.org/Vol-3235/paper17.pdf.

(PW5) Fostering trust with transparency in the data economy era: An integrated ethical, legal, and knowledge engineering approach. (2022) **B. Esteves**, H. Asgarinia, A. Chomczyk Penedo, B. Mutiro, D. Lewis. In *Proceedings of the 1st International Workshop on Data Economy*, pp. 57-–63. https://doi.org/10.1145/3565011.3569061.

(PW6) Towards an Architecture for Data Altruism in Solid. (2023) **B. Esteves**. In *22nd International Semantic Web Conference: Posters, Demos, and Industry Tracks*. https://ceur-ws.org/Vol-3632/ISWC2023_paper_491.pdf

(PW7) Using Patterns to Manage Governance of Solid Apps. (2023) **B. Esteves**, H. J. Pandit. In *14th Workshop on Ontology Design and Patterns (WOP 2023@ISWC 2023)*. https://ceur-ws.org/Vol-3636/paper5.pdf.

1.4.4 Oral presentations

The following presentations were given during the realisation of this Thesis.

(OP1) Can privacy terms be negotiated in Solid's personal datastores? **B. Esteves**. Short talk at the *2021 IEEE Symposium and Workshops on Security & Privacy* (26/May/2021).

(OP2) 'Who should I trust with my data?': are decentralised technologies the answer to achieving ethical and lawful data governance practices? H. Asgarinia, A. Chomczyk Penedo, **B. Esteves**, D. Lewis, B. Mutiro. Presentation at the *Data and the Common Workshop 2022* (04/March/2022).

(OP3) Access control policies for Solid. **B. Esteves**. Demonstration at the *2022 COST EU Workshop on Privacy Issues in Distributed Social Knowledge Graphs* (14/June/2022).

(OP4) Establishing Data (Re-)Use Agreements Through Semantic Policies for Legally-aware Data Sharing. **B. Esteves**. Lightning talk at the *SEMIC Conference 2022* (06/December/2022).

(OP5) Altruistic (Re-)Use of Health Data through Semantic Policies. **B. Esteves**. Short talk at the *2nd DPSN International Data Protection Day* (27/January/2023).

(OP6) Privacy Receipts in Solid Pods. **B. Esteves**, J. Lindquist. Presentation at the *2023 COST EU Workshop on Privacy Issues in Distributed Social Knowledge Graphs* (13/February/2023).

(OP7) Policies in Solid: The Road Ahead. **B. Esteves**. Presentation at the *Solid Symposium 2023* (31/March/2023).

(OP8) Enhancing Solid with Legally-aware Policies. **B. Esteves**. Presentation at the *Governing Artificial Intelligence International Symposium 2023* (23/May/2023).

1.5 Projects

The following projects funded the work presented in this Thesis:

PROTECT ITN: Protecting Personal Data Amidst Big Data Innovation (PROTECT) is an EU-funded Innovative Training Network project with the goal of developing *"new ways of empowering users of digital services to understand the risks they take when they go online and to offer new ways to enable companies to incorporate data protection into digital services"* and train *"a new generation of 14 early stage researchers who will integrate and apply arguments, analyses, and tools from across the fields of law, ethics and knowledge engineering"*[6]. This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 813497.

[6]Extracted from https://cordis.europa.eu/project/id/813497 (accessed on 15 March 2024).

AURORA: Achieving a new European Energy Awareness (AURORA) is an EU-funded Innovation Action project whose main objective is to *"empower several thousand citizens across five locations in Denmark, England, Portugal, Slovenia, and Spain to make more informed energy decisions"*[7]. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 101036418.

INESData: Infraestructura para la Investigación de Espacios de Datos (INESData) is an EU-funded project with the main goal of creating and installing a data governance structure and technological components for common data spaces. This project has received funding from the European Union's NextGenerationEU funding programme.

COST DKG: COST Action on Distributed Knowledge Graphs (DKG), with grant agreement No. CA19134, whose main goal is to *"create a research community for deployable Distributed Knowledge Graph technologies that are standards-based, and open, embrace the FAIR principles, allow for access control and privacy protection, and enable the decentralised publishing of high quality data"*[8].

1.6 Research stays

The research stays done in the context of this Thesis are outlined below.

01/10/2021 – 01/12/2021 (2 months): Research stay at *EmPushy, Dublin, Ireland*[9], supervised by Dr. Kieran Fraser. During this stay, the Annotation of Push-Notifications (APN) ontology was created to annotate push-notification datasets and train models to identify the presence of personal data in notifications' text, the intent of the notification, its persuasiveness, and so on. As this ontology is out of the scope of this Thesis, its description is omitted from this document. This work resulted in the publication of two conference papers, (PC1) and (PC2) [Esteves et al., 2022b,c]. An analysis of which data to track in EmPushy's tools, and respective GDPR requirements to fulfil, was also performed with the EmPushy team. This stay was funded by the PROTECT ITN.

01/02/2022 – 31/07/2022 (6 months – half-time): Virtual research stay at *Inrupt, Inc., Boston, United States of America*[10], supervised by Pat McBennett and Nicolas Mondada. During this stay, an overview of relevant vocabularies related to the Solid ecosystem was performed. Moreover, this Thesis work on the ODRL profile for Access Control (OAC) was improved with the requirements brought by Inrupt's use cases and a Solid application (Solid ODRL access control Policies Editor – SOPE) was developed to generate and store OAC policies in Solid Pods. This work resulted in the publication of the (PW3) workshop paper [Esteves et al., 2022e]. This stay was funded by the PROTECT ITN.

[7] Extracted from https://cordis.europa.eu/project/id/101036418 (accessed on 15 March 2024).
[8] Extracted from https://www.cost.eu/actions/CA19134/ (accessed on 15 March 2024).
[9] https://www.empushy.com/ (accessed on 15 March 2024)
[10] https://www.inrupt.com/ (accessed on 15 March 2024)

01/09/2022 – 01/12/2022 (3 months): Research stay at *ADAPT Centre, Trinity College Dublin, Dublin, Ireland*[11], supervised by Prof. Dr. Harshvardhan J. Pandit and Prof. Dr. Dave Lewis. During this stay, the Policy LAnguage for Solid's Metadata-based Access control (PLASMA) was developed. We also contributed to the development of the Data Privacy Vocabulary (DPV) specifications, including writing documentation and use cases, in particular, related to the exercising of data subjects' rights and the new DGA law. This work resulted in the publication of one conference and two workshop papers, (PC4), (PW6), and (PW7) [Esteves et al., 2023, Esteves, 2023, Esteves and Pandit, 2023]. This stay was funded by the PROTECT ITN.

01/03/2023 – 18/03/2023 (3 weeks): Short-Term Scientific Mission at *KNoWS, IDLab, Ghent University, Ghent, Belgium*[12], supervised by Prof. Dr. Ruben Verborgh. The main objective of this stay was to discuss and establish technical and legal requirements to align Solid with data protection principles and understand current issues and solutions that need to be dealt with and reused to implement such requirements in decentralised data-sharing environments. This mission was funded by the DKG COST Action.

[11] https://www.adaptcentre.ie/ (accessed on 15 March 2024)
[12] https://knows.idlab.ugent.be/ (accessed on 15 March 2024)

Chapter 2

State of the Art

The content of this Chapter has already been partially included in the articles published during this Thesis [Esteves et al., 2021, Esteves and Rodríguez-Doncel, 2022a, Asgarinia et al., 2023, Esteves and Pandit, 2023, Florea and Esteves, 2023].

This Chapter presents the state of the art on the representation of policies and personal data processing metadata in the context of determining access to decentralised personal data systems, focusing on solutions that use Semantic Web technologies and cater to data protection law requirements. Since there are different areas being covered in this state of the art analysis, each topic will be first introduced with a list of criteria used to perform said analysis. The prefixes and namespaces used in the Listings in this Chapter are explicitly defined in the Namespaces list.

Thus, a literature review was performed in the subsequent areas, following the methodology described in Section 3.7.1:

2.1 Decentralising the access to personal data with Solid

2.2 Representing personal data processing information

2.3 Using policy languages to specify access control conditions

Through this analysis, a series of gaps and challenges in the representation of privacy terms was identified, in order to have a legally-aligned Solid environment, and is described in Section 2.4.

2.1 Decentralising the access to personal data with Solid

As attested in Section 1.3.6, current efforts are underway to decentralise today's Web. By decoupling data from applications, Web users will have their data stored in an environment they can control and can choose what data they want to make publicly accessible or accessible only to certain users while using their application of choice to manage said data. This represents a significant paradigm shift regarding users' current online experience – instead of being locked away in Big Tech companies' storage servers, data can be stored on individual personal datastores maintained by a provider chosen by the user or hosted by the user itself on its private server. In this context,

the Solid project [Sambra et al., 2016, Mansour et al., 2016] has been gaining prominence as it relies on Web standards to achieve this degree of decentralisation – its ultimate goal is to give Web users a personal datastore, i.e., a Pod, per user, with a granular access control system managed by the user, which they can use to select which people and/or applications have access to the resources stored on their Pod. As such, with Solid, applications and the companies behind them do not store the (personal) data of their users, acting only as interfaces that can read, write, or append data to/from Pods. Such an ecosystem *"fosters innovation and competition through separate markets for data and applications"* [Verborgh, 2017], while allowing Web users to exert a degree of control over their data that is currently impossible to wield.

In particular, the Solid protocol [Capadisli et al., 2022] describes how servers and apps should behave by relying on the following Web standards:

- HTTP [Fielding et al., 2022] – The Hypertext Transfer Protocol describes an architecture and semantics for *"distributed, collaborative, hypertext information systems"* to share data.
- RDF [Cyganiak et al., 2014] – The Resource Description Framework defines a data model to represent information in the Web, including a schema [Brickley and Guha, 2014] and serialisation syntaxes for storing and exchanging RDF such as Turtle [Prud'hommeaux and Carothers, 2014] and JSON-LD [Gregg Kellogg et al., 2020].
- LDP [Speicher et al., 2015] – The Linked Data Platform specification expresses how to use *"HTTP for accessing, updating, creating and deleting resources from servers that expose their resources as Linked Data"*.
- SPARQL [Harris and Seaborne, 2013] – The SPARQL language can be used to query RDF databases.
- WebID-TLS [Story et al., 2014] – A protocol that uses WebIDs to authenticate users on the Web.
- OIDC [Sakimura et al., 2014] – The OpenID Connect standard is an authentication protocol to assert the user's identity.

Moreover, Table 2.1 provides an overview of Solid-related concepts, their definitions, and corresponding classes and properties already modelled in Solid-related vocabularies. The authentication and authorisation protocols compose Solid's two main building blocks – an up-to-date list of Solid specifications, including technical reports for both aforementioned building blocks, is maintained by the Solid Community Group at `https://solidproject.org/TR/`. Authentication is a necessary feature to identify users when they want to log into their Pod and/or when they want to use an app to perform a certain action over resources stored in their Pod. Thus, Solid's authentication protocol uses Solid's WebID specification to identify agents through URIs, as specified in Table 2.1, which when dereferenced return a WebID profile document that should include information regarding the identity provider chosen by the Solid user and the Pod storage location and may include information regarding an available inbox where users and applications can leave messages to the user [Balseiro et al., 2022]. In addition, to verify the identity of agents, the Solid Protocol recommends the usage of the Solid OIDC protocol[1] [Coburn et al., 2022], however

[1]A Solid-OIDC Primer [Morgan et al., 2022] is also being developed to provide additional knowledge on Solid OIDC's authentication flows.

additional authentication methods, such as the previously mentioned WebID-TLS, can also be implemented.

The authorisation protocol specifies mechanisms used by Solid servers to reply to requests of particular users or apps to have access to certain resources, containers of resources, or even to the whole Pod. Furthermore, the Solid protocol states that a Solid server *"MUST conform to either or both Web Access Control (WAC) and Access Control Policy (ACP) specifications"* in order for it to be a compliant Solid server. Specific authorisation use cases and requirements are documented by the Solid Editorial Team [2023] and further details on the authorisation methods will be given in Section 2.1.1. In addition to the authorisation specifications, there is a third protocol being developed to ensure data interoperability and (re)usability across Pod providers, agents, and applications – the Solid Application Interoperability (SAI) specification [Bingham et al., 2023].

Table 2.1: Overview of Solid-related concepts, their definitions, and related terms modelled on Solid specifications.

Concept	Definition	Solid vocabularies
Pod	A personal datastore that conforms to the Solid protocol	`pim:Storage`
Resource	Target asset stored in a Pod identified by a URI. Container resources can contain other resources including containers	`acl:accessTo`, `acp:target`
Inbox	Container resource for messages sent to an agent	`ldp:inbox`, `interop:hasInbox`
Server	Server capable of hosting resources and responding to resource requests	
App	An application that reads and writes data to Pods	`interop:Application`, `acp:client`, `acl:origin`
Agent	A person, social or virtual entity identified by a URI	`interop:Agent`, `acl:agent`, `acp:agent`
Pod Owner	Agent that has control over all resources in a Pod including access control resources	`solid:owner`, `acp:owner`
WebID	URI that acts as a primary identifier for agents, which, when dereferenced, resolves to an identity profile document (WebID profile)	
Identity Provider	Entity implementing the identity service capable of authenticating a WebID	`solid:oidcIssuer`, `acp:issuer`
Pod Provider	Entity providing the storage space and maintaining the server implementation	
Policy	Conditions for accessing the Pod and its resources	`acp:Policy`
Registry	Records where agents can store and find different types of data for different purposes	`interop:Registry`

2.1.1 Access control and interoperability in Solid

As discussed in the previous Section, there are two distinct access control methods being specified in the Solid ecosystem – both use URIs to identify resources and users, while WAC [Capadisli, 2022] relies on ACLs and ACP [Bosquet, 2022] on Access Control Resources (ACRs) to specify who is authorised or refused access and access grants to represent the final authorisation decision. While server providers can implement only one of the authorisation protocols, Solid applications can not do the same or else they will not work with server providers that use a distinct protocol from the one they choose to implement. Listings 2.1 and 2.2 provide examples of both types of access control statements. As is visible by the examples, both solutions do not have the depth to deal with the users *'Right to be Informed'* (Arts. 13 and 14) [2016b], since these models do not contain the terms to specify the purpose for accessing data on Pods, the personal data categories being consulted, used legal basis or even information on the identity of application developers.

Listing 2.1 WAC authorisation that makes a WebID profile, https://solidweb.me/besteves4/profile/card, readable by any agent.

```
<#public> a acl:Authorization ;
    acl:agentClass foaf:Agent ;
    acl:accessTo <https://solidweb.me/besteves4/profile/card> ;
    acl:mode acl:Read .
```

Listing 2.2 ACP authorisation that makes a WebID profile, https://solidweb.me/besteves4/profile/card, issued by https://solidweb.me/, readable by any agent using any application.

```
<#public> a acp:AccessGrant ;
    acp:grant acl:Read ;
    acp:context [
        acp:agent acp:PublicAgent ;
        acp:target <https://solidweb.me/besteves4/profile/card> ;
        acp:client acp:PublicClient ;
        acp:issuer <https://solidweb.me/> ] .
```

Moreover, these access protocols were deemed not enough to ensure the interoperability of agents, data, and applications, and as such an interoperability specification is being developed to describe the implementation of agent, data, and access registries, to track user interactions with other agents, to keep records of where data is being stored and to manage access grants given to other agents [Bingham et al., 2023]. Listing 2.3 provides an example of a registry set that should only be readable by the Pod owner and Listing 2.4 an example of an authorisation registry, which contains an `AccessAuthorization` for the `projectron` app which requires data with a particular shape[2]. While SAI is a step forward in Solid towards having a more transparent ecosystem, it is still in the early stages of development, and as such it is still not clear how this specification will fit in with the existing access control protocols or how it is going to be implemented/enforced. In

[2]SAI assumes the use of ACL access modes which are still not approved, e.g., `acl:Create`, `acl:Update`, `acl:Delete`, and are under discussion on the Solid CG authorisation panel (see the issue at https://github.com/solid/authorization-panel/issues/253).

addition, as is illustrated by Listing 2.4, SAI does not entirely fulfil GDPR requirements, e.g., it does not provide transparency regarding the purpose for access or which type of data is being accessed nor does it provide information regarding the identity of the entities that develop/provide the apps.

Listing 2.3 Registry set, established according to the SAI specification, that stores private information regarding the storage location of registries of https://solidweb.me/besteves4/.

```
PREFIX beatriz-registry: <https://solidweb.me/besteves4/registry/>
PREFIX beatrizWork-registry:
 ↪ <https://solidweb.me/besteves4-work/registry/>

beatriz-registry: a interop:RegistrySet ;
  interop:hasAgentRegistry beatriz-registry:agents ;
  interop:hasAuthorizationRegistry beatriz-registry:authz ;
  interop:hasDataRegistry beatriz-registry:data ,
   ↪ beatrizWork-registry:data .
```

Furthermore, the idea of having registries of data and applications is also compatible with the graph-centric interpretation of a Pod debated by Dedecker et al. [2022]. Certain apps might require the presence of particular data stored in a particular container – this will cause an interoperability problem as it is something that cannot be standardised across the ecosystem for all apps. With a graph-centric approach, *"each Solid pod is a hybrid, contextualized knowledge graph, wherein 'hybrid' indicates first-class support for both documents and RDF statements, and 'contextualized' the ability to associate each of its individual documents and statements with metadata such as policies, provenance, and trust"*. With such metadata, including context and provenance metadata, distinct views of the Pod can be rendered as required by different applications or agents. Moreover, data request policies can simply be appended to the *'Pod as a Graph'*, without the need to have it hard-coded in the app, and can be viewed by users using graphic-centric Solid apps.

2.1.2 Solid and data protection

Only recently has the debate on data protection reached the concerns of Solid's developers, mainly with regard to issues of control and privacy of personal data. In addition, beyond the access control mechanisms discussed in the previous Section, there are also researchers starting to work on *'usage control'*, a process which has as its main concern the enforcement of the users' policies after the access to the data has already been given [Akaichi, 2022, Havur et al., 2020]. As such, in this Section, we describe the existing body of work on data protection and governance aspects of Solid-related technologies, with a particular focus on GDPR-related academic and industrial research.

Exercising of data subject rights De Mulder et al. [2021] developed PROV4ITDaTa[3], a configurable application that facilitates the exercising of the data subject's *'right to data portability'* (Art. 20) [2016b], using open sources resources such as RML.io[4] [Dimou et al., 2014] – to access and

[3]The source code is available at https://github.com/RMLio/prov4itdata-web-app, under an MIT license (accessed on 14 August 2023).

[4]https://rml.io/ (accessed on 14 August 2023)

Listing 2.4 Authorisation registry of https://solidweb.me/besteves4/.

```
PREFIX beatriz-authz: <https://solidweb.me/besteves4/registry/authz/>
PREFIX projectron: <https://projectron.app/>
PREFIX projectron-shapetrees: <https://projectron.app/shapetrees/>

beatriz-registry:authz a interop:AuthorizationRegistry;
    interop:hasAccessAuthorization beatriz-authz:projectron .

beatriz-authz:projectron a interop:AccessAuthorization ;
    interop:grantedBy <https://solidweb.me/besteves4/profile/card#me> ;
    interop:grantedWith <https://authz.agent/id> ;
    interop:grantedAt "2023-07-31T11:53:01Z"^^xsd:dateTime ;
    interop:grantee projectron:id ;
    interop:hasAccessNeedGroup projectron:need-group ;
    interop:hasDataAuthorization beatriz-authz:54a1b6a0 .

projectron:need-group a interop:AccessNeedGroup ;
    interop:accessNecessity interop:accessRequired ;
    interop:accessScenario interop:PersonalAccess ;
    interop:authenticatesAs interop:SocialAgent ;
    interop:hasAccessDescriptionSet projectron:access-en ;
    interop:hasAccessNeed projectron:need-project .

projectron:need-project a interop:AccessNeed ;
    interop:registeredShapeTree projectron-shapetrees:ProjectTree ;
    interop:accessNecessity interop:accessRequired ;
    interop:accessMode acl:Read, acl:Create ;
    interop:creatorAccessMode acl:Update, acl:Delete .

beatriz-authz:54a1b6a0 a interop:DataAuthorization ;
    interop:grantee projectron:id ;
    interop:registeredShapeTree projectron-shapetrees:ProjectTree ;
    interop:accessMode acl:Read, acl:Create ;
    interop:creatorAccessMode acl:Update, acl:Delete ;
    interop:scopeOfAuthorization interop:All ;
    interop:satisfiesAccessNeed projectron:need-project .
```

generate interoperable Linked Data datasets using the Schema.org or DCAT vocabularies, Solid – to store the datasets, and Comunica[5] [Taelman et al., 2018] – to query the datasets, and promotes transparency by automatically generating and recording provenance metadata using the PROV Ontology standard. The PDS Interop collaboration [2021d], an effort that started with the goal to make Solid and Nextcloud[6] interoperable, also developed an app, the Solid Migrator App[7], to assist in the migration of Pod resources to a different Pod, independently of the Pod provider.

GDPR principles Pandit [2023] describes Solid as a *'cloud technology'*, according to ISO standards, provides a theoretical discussion on how GDPR principles apply to Solid and suggests how to extend its specifications to deal with such requirements. Esposito et al. [2023] also provide a theoretical analysis of technical security and privacy measures to assist Solid developers in complying with the GDPR – a mapping of Solid actors and respective legal roles is provided for accountability, as well as security measures to ensure data confidentiality and minimisation and protocols to safeguard the data subjects' rights to be notified (Art. 19) [2016b], to object (Art. 21) [2016b] and to not be a target of automated decision-making (Art. 22) [2016b]. Van Damme et al. [2022] present a qualitative analysis of a series of plenary sessions with academia, governments, citizens, and industry regarding the adoption of decentralised personal datastore technologies. The main challenges that were identified are related to social, technical, legal, and ecosystem issues, which need to be considered for the *"development of an interdisciplinary research agenda"*. In terms of legal challenges, the core aspects that were discussed are related to control, portability, compliance, accountability, delegation of consent, and the usage of other legal bases such as legitimate interests. The role of intermediates, promoted by the DGA, was also discussed. Digita[8], a Belgian startup that offers Solid-based identity and storage solutions, published a research report reflecting on accountability aspects related to the implementation of Solid products, mainly regarding the lawfulness of data usage and transfer to recipients, particularly based on consent, and the specificity and compatibility of purposes [De Bot and Haegemans, 2021]. Bailly et al. [2023] propose to use the SAI and DPV vocabularies to specify access and usage control policies, respectively, and provide a prototype User Interface (UI) for users to consent to data requests[9], which the authors found to have a low score in terms of usability.

Domain-specific use cases Several health-related use cases have been developed by the Solid research community. Among them, TIDAL[10] (ciTIzen-centric DAta pLatform), a Solid-powered application, has been developed by Sun et al. [2023] for healthcare researchers to request consent from citizens to use their data for health-related research. DPV is used to limit the purpose for which the data can be used, DPV-PD and other health-related vocabularies to restrict the categories of personal data, and privacy-preserving data analysis algorithms to preserve data confidentiality.

[5] https://comunica.dev/ (accessed on 14 August 2023)

[6] https://nextcloud.com/ (accessed on 20 August 2023)

[7] The source code is available at https://github.com/pdsinterop/solid-migrator-app, under an MIT license (accessed on 20 August 2023).

[8] https://www.digita.ai/ (accessed on 15 August 2023)

[9] The source code is available at https://github.com/HBailly/solid-auth-ui/, under a GNU General Public License v3.0 (accessed on 15 August 2023).

[10] The source code is available at https://github.com/sunchang0124/TIDAL, under an MIT license (accessed on 15 August 2023).

Janeiro Digital[11] [2021c] is working with the United Kingdom's National Health Service (NHS) to manage and use patient data from several systems, providing patients with individual Solid Pods and giving healthcare professionals access to data through the Solid protocol. Ammar et al. [2020, 2021] discuss the implementation of a *'Personal Health Library'* using Solid *"to deliver tailored push notifications to support behavior change related to chronic disease self-care"* based on sensor readings and other information. In addition, they are developing an app to allow users to decide what data should be stored in the Pod, and who should have access to it, and to share their data with other research initiatives.

In addition to the work of De Bot and Haegemans [2021], Buyle et al. [2020] also focus on government-related use cases. In this work, a Solid app for Flemish citizens, that allows them to share data with government administrations and to reuse said data in different contexts, is described to increase the accuracy of personal data which is difficult to keep updated, e.g., telephone numbers or email addresses, and to allow data portability. Wang's thesis also discusses the usage of Solid to enhance governmental services, including provisions to improve the access (Art. 15) [2016b] and rectification (Art. 16) [2016b] rights exercised in this context, including a use case scenario of applying to a social house and dealing with the subsequent changes to the address.

Karamel[12] also partnered with Digita to create a human resources platform for applicants and recruiters to find new jobs [Verstraete et al., 2022] – users can manage data stored in Solid Pods through an app that allows applicants to revoke access and to request to be forgotten (Art. 17) [2016b] by recruiters.

Tóth [2022] developed a prototype architecture for the domain of hospitality where users can use Solid applications to book/manage accommodations and edit personal information. This proof of concept[13] allows its users to request deletion (Art. 17) [2016b] and rectification (Art. 16) [2016b] of data and of copies of said data.

Van de Wynckel and Signer [2022] are researching the usage of Solid to develop transparent indoor positioning systems that store individual and sensor data in Solid Pods in an interoperable format. In addition, they developed an app that reads the user's personal position, orientation, and velocity from the Pod and displays them along with additional information[14].

The described solutions can be compared through Figure 2.1. Each solution was analysed in terms of whether it assists in the exercising of data subject rights or the implementation of a certain GDPR principle, as well as the type of solution developed by the authors of the paper. Works describing Solid apps are marked with a **black** shape, identity provider solutions with a **orange** shape, Pod provider solutions with a **blue** shape and theoretical work with a **red** shape. In addition, works marked with a ○ discuss a government-related use case, △ a health-related use case, ◇ a hospitality-related use case, ☆ a location-related use case, □ a human resources-related use case and **X** represents work without a particular domain-related use case.

[11] https://www.janeirodigital.com/ (accessed on 15 August 2023)
[12] https://karamel.career/ (accessed on 20 August 2023)
[13] The source code is available at https://github.com/gergelyth/solid-hotel, under an MIT license (accessed on 21 August 2023).
[14] The source code is available at https://github.com/OpenHPS/ipin2022-solid/ (accessed on 23 August 2023).

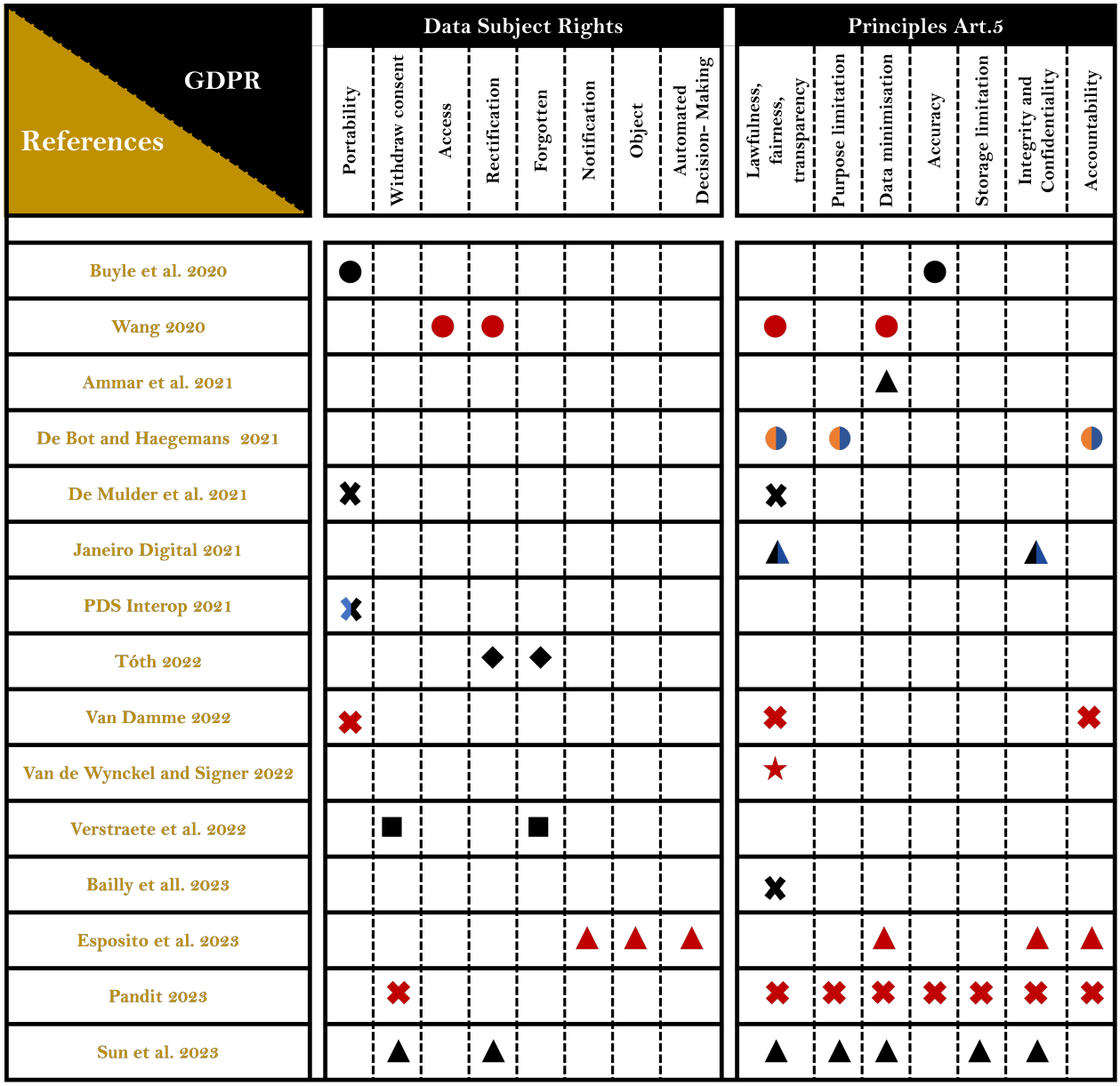


Figure 2.1: Comparison of existing work on Solid and data protection topics. Each work was analysed in terms of whether it assists in the exercising of data subject rights or in the implementation of a certain GDPR principle, as well as the type of solution they developed – works describing Solid apps are marked with a **black** shape, identity provider solutions with a **orange** shape, Pod provider solutions with a **blue** shape and theoretical work with a **red** shape. In addition, works marked with a ○ discuss a government-related use case, △ a health-related use case, ◇ a hospitality-related use case, ☆ a location-related use case, □ a human resources-related use case and **X** represents work without a particular domain-related use case.

As illustrated by the Figure, most works concentrate on fulfilling one or more GDPR principles, with a strong focus on the *'lawfulness, fairness and transparency'* principle. Distinct works were also found to tackle the right to data portability, withdrawal of consent, and rectification, while no specific work was found on the right to restrict the processing of personal data.

2.1.3 Other decentralised technologies

In addition to Solid and its stack of technologies, there is a set of tools and resources that support decentralisation and are either being discussed or should be discussed to be included in the Solid ecosystem. We briefly present them in this Section as they should be considered in future research in this area.

Decentralised Identity The Decentralized identifiers (DIDs) data model is a recent W3C Recommendation that enables *"individuals and organizations to generate their own identifiers using systems they trust"* [Sporny et al., 2022][15]. In contrast with centralised settings, DIDs allow something or someone to be identified by a globally unique identifier which is detached from centralised registries, identity providers, or certificate authorities. Moreover, DIDs work in a similar fashion to WebIDs – DIDs are still URIs that can be dereferenced to return a DID document which *"can express cryptographic material, verification methods, or services [...] to prove control of the DID"*. The European Commission is also promoting the emergence of digital identity wallets for EU citizens, residents, and businesses to identify themselves, both online and offline, and to exchange certain types of personal identification data such as birth certificates or driving licenses. In this context, the electronic IDentification, Authentication and trust Services (eIDAS) regulation [2014], and its amendment [2021b], puts forward *"rules for trust services"* and *"a legal framework for electronic signatures, electronic seals, electronic time stamps, electronic documents, electronic registered delivery services and certificate services for website authentication"*, while the 2021 amendment explicitly adds *"conditions for the issuing of European Digital Identity Wallets"* and additional considerations derived from the GDPR.

Verifiable Credentials W3C's Verifiable Credentials (VCs) data model specification describes *"a standard way to express credentials on the Web in a way that is cryptographically secure, privacy-respecting, and machine-verifiable"* [Sporny et al., 2023], using technologies such as digital signatures. As in the physical world, VCs can be used to identify data subjects, to provide government-issued documents, e.g., ID cards or passports, or to deliver information on how the credential was created/derived and other constraints such as validity period or conditions for use. An example of a VC, with a DID-identifiable subject, is presented in Listing 2.5. The usage of VCs is being contemplated in the ACP specification, however, no specific details are provided regarding its implementation/development.

Moreover, Braun and Käfer [2022b,a] have previously published work on using VCs and RDF-

[15] A DID Solid method specification is in the early stages of development, with no official implementations being known to date – available at `https://solid.github.io/did-method-solid/` (accessed on 19 August 2023).

Listing 2.5 Verifiable credential with terms of use where the issuer is prohibiting verifiers from archiving a degree credential, extracted from Sporny et al. [2023], with the credential subject being identified by a DID.

```
{
    "@context": [
        "https://www.w3.org/ns/credentials/v2",
        "https://www.w3.org/ns/credentials/examples/v2"
    ],
    "id": "http://university.example/credentials/3732",
    "type": ["VerifiableCredential", "ExampleDegreeCredential"],
    "issuer": "https://university.example/issuers/14",
    "validFrom": "2010-01-01T19:23:24Z",
    "credentialSubject": {
        "id": "did:example:ebfeb1f712ebc6f1c276e12ec21",
        "degree": {
          "type": "ExampleBachelorDegree",
          "name": "Bachelor of Science and Arts"
        }
    },
    "termsOfUse": [{
        "type": "IssuerPolicy",
        "id": "http://example.com/policies/credential/4",
        "profile": "http://example.com/profiles/credential",
        "prohibition": [{
            "assigner": "https://university.example/issuers/14",
            "assignee": "AllVerifiers",
            "target": "http://university.example/credentials/3732",
            "action": ["Archival"]
        }]
    }]
}
```

star[16] to sign and verify Web resources using a Solid app[17] and on using VCs and Linked Data Notifications (LDNs)[18], extending previous work from Ezike, to request attribute-based access to Solid Pods[19].

2.2 Representing personal data processing information

The Web of Data and its semantic specifications are thriving, with the W3C guiding this effort to have machine-readable and interoperable linked data on the Web, described by open standards that

[16] RDF-star[Arndt et al., 2023] is an under-development W3C specification that *"extends RDF with a convenient way to make statements about other statements"*.

[17] The source code is available at https://github.com/uvdsl/solid-web-ldsig, under an MIT license (accessed on 19 August 2023).

[18] The W3C LDNs Recommendation [Capadisli and Guy, 2017] is a protocol that describes how servers can send and retrieve RDF-based messages, sent/retrieved by applications.

[19] The source code is available at https://github.com/uvdsl/solid-vc-pwa/, under an MIT license (accessed on 19 August 2023).

promote portability and extensibility, allow for seamless integration of data from different origins and re-usage across distinct Web applications and services [Berners-Lee et al., 2001]. Accordingly, a wave of vocabularies and ontologies has appeared in the last few years to formalise common concepts, such as objects or entities, or terms from specific domains, such as legal or medical ontologies. More recently, with the increasing concerns around personal data processing abuse by Big Tech companies, data protection has been on the agenda of governments in a worldwide manner, with the EU taking the lead with its data protection law, the GDPR. This has led to a new wave of personal data protection law ontologies being developed to help companies comply with GDPR's legal requirements and which also intend to assist individuals in managing their personal data.

In the following Sections, the criteria used to analyse each specified data protection vocabulary are included, as well as a description of each identified solution.

2.2.1 Criteria for analysis

A description of the core terms formalised in the ontology, as well as dependencies on other existing works, is provided for each identified solution, and, when available, information on case studies where it has been applied. As for the analysis of their representational abilities, ontologies are evaluated in terms of what privacy terms mentioned in GDPR's data subject rights they can represent (data subject rights are detailed in Section 1.3.4), to assess to what extent they can be used by data subjects to exercise their rights and by data controllers to manage compliance.

For such rights to be exercised by data subjects and fulfilled by data controllers, a set of privacy terms must be modelled, namely the terms identified in Table 2.2, which is derived from Esteves and Rodríguez-Doncel [2022a]. These terms will be used to compare the described privacy and data protection vocabularies and ontologies and identify representational gaps in the existing solutions. The outcomes of this comparative analysis will be provided in Section 2.2.3.

2.2.2 Personal data protection vocabularies

In this Section, a systematic description of each identified work is provided, including an example that uses the vocabulary's concepts, when available. Moreover, Table 2.3 provides an overview of these solutions and supplies information about the creators of the resources, their versions, the date of publication, and the date of the last known update. Said solutions are then analysed in chronological order regarding the date of publication and a dependency graph – a chart that demonstrates the relations between the reviewed vocabularies and its dependencies and extensions – is presented in Figure 2.2.

To complement the description of vocabularies presented in this Section, additional documentation and resources were published in a Web page[20], including diagrams and code examples.

[20] Available at https://w3id.org/people/besteves/phd/sota/ontologies. Its public repository can be consulted at https://w3id.org/people/besteves/phd/sota/repo for further improvement when new solutions appear.

Table 2.2: Privacy terms to be represented and respective identifiers (I*). The GDPR articles that mention these terms are also specified.

I*	Informational Items	**GDPR Article(s)**
I1	Controller identity	**13.1(a), 14.1(a)**
I2	Controller contact details	**13.1(a), 14.1(a)**
I3	Controller's representative identity	**13.1(a), 14.1(a)**
I4	Controller's representative contact details	**13.1(a), 14.1(a)**
I5	DPO contact details	**13.1(b), 14.1(b)**
I6	Purposes of the processing	**13.1(c), 14.1(c), 15.1(a)**
I7	Legal basis of the processing	**6.1, 9.2, 13.1(c), 14.1(c)**
I8	Legitimate interests	**6.1(f), 13.1(d), 14.2(b)**
I9	Recipients / categories of recipients	**13.1(e), 14.1(e), 15.1(c), 17.2, 19**
I10	Transfers to third countries	**13.1(f), 14.1(f)**
I11	Retention period	**13.2(a), 14.2(a), 15.1(d)**
I12	Data subject's rights	**13.2(b), 14.2(c), 15.1(e)**
I13	Right to withdraw consent	**6.1(a), 9.2(a), 13.2(c), 14.2(d)**
I14	Right to lodge a complaint	**13.2(d), 14.2(e), 15.1(f)**
I15	Statutory or contractual obligation details	**13.2(e)**
I16	Existence of automated decision making	**13.2(f), 14.2(g), 15.1(h), 22.1, 22.4**
I17	Categories of personal data	**9.1, 14.1(d), 15.1(b)**
I18	Source of personal data	**14.2(f), 15.1(g)**
I19	Grounds to not comply with information right	**13.4, 14.5**
I20	Safeguards related to the transfer to a third country	**15.2**
I21	Copy of personal data	**15.3, 20.1**
I22	Request to complete incomplete personal data	**16**
I23	Grounds to request erasure of data	**17.1**
I24	Technical measures taken to erase data	**17.2**
I25	Recipients contact details	**17.2, 19**
I26	Grounds to not comply with right of erasure	**17.3**
I27	Grounds to request restriction of processing	**18.1**
I28	Transfer data directly between controllers	**20.2**
I29	Grounds to not comply with right to object	**21**
I30	Grounds to not comply with right not to be subjected to decision making	**22.2**

Table 2.3: Brief description of the vocabularies described in Section 2.2.2.

Abbreviation (Section)	Full Name	Creators	Version	Date of publication	Last update
DPKO, DPRO (2.2.2)	Data Protection Knowledge Ontology, Data Protection Reasoning Ontology	Casellas et al.	-	2008	2010
DPO (2.2.2)	Data Protection Ontology	Bartolini and Muthuri	-	2015	2016
GDPRov (2.2.2)	GDPR Provenance Ontology	Pandit and Lewis	0.7	2017	2019
Cloud (2.2.2)	Cloud GDPR ontology	Elluri and Joshi	-	2018	-
PrOnto (2.2.2)	Privacy Ontology for legal reasoning	Palmirani et al.	-	2018	-
GConsent (2.2.2)	GDPR Consent ontology	Pandit et al.	0.5	2018	-
IMO (2.2.2)	Information Model Ontology	Lioudakis and Cascone	1.0	2018	-
DPV (2.2.2)	Data Privacy Vocabulary	Pandit et al.	1.0	2018	2022
GDPRtEXT (2.2.2)	GDPR text EXTensions	Pandit et al.	0.7	2018	2020

NEURONA

The NEURONA project [Casellas et al., 2010], developed by S21SEC[21] and IDT-UAB[22], created two ontologies based on the pre-GDPR Spanish personal data protection regulation [2008b] – the Data Protection Knowledge Ontology (DPKO) and the Data Protection Reasoning Ontology (DPRO). These ontologies are, however, not publicly available for re-usage.

The main objectives of the ontologies developed in the context of this project were to represent security measures for files containing personal data and reason over their correctness. DPKO's main classes are **data**, **consent**, **purpose**, **person**, **security measures** and **security degree**. In relation to the data class, categories such as health data are defined and associated with special security measures. The consent should be given by the data subject in an unambiguous way and for a specific purpose. In addition, technical and organisational measures (TOMs) for data security, such as access control measures or authentication procedures, are modelled and connected to the nature of the data, taking into consideration the security level associated with the type of data or how the data was obtained. For example, a file with health data obtained without consent should have high-level security measures, while a file with anonymised data can implement low-level measures. DPRO is then used to access data protection compliance by reasoning over the measures applied to files.

DPO

The Data Protection Ontology (DPO) [Bartolini and Muthuri, 2015, Bartolini et al., 2017] concentrates on the modelling of data protection principles and data controller obligations. It is based on an early version of the GDPR, prior to its implementation in May 2018, on the Data Protection Directive [1995] and the *Handbook on European data protection law* [European Union Agency for Fundamental Rights and Council of Europe, 2018] and reuses concepts from the LKIF-Core [Hoekstra et al., 2007] ontology.

The core classes of DPO are **data protection principles**, **rules** for processing and transferring data, and **data subjects rights**. The ontology is designed so that each rule and right is linked to at

[21] https://www.s21sec.com/ (accessed on 16 July 2023)

[22] https://portalrecerca.uab.cat/en/organisations/law-and-technology-institute (accessed on 16 July 2023)

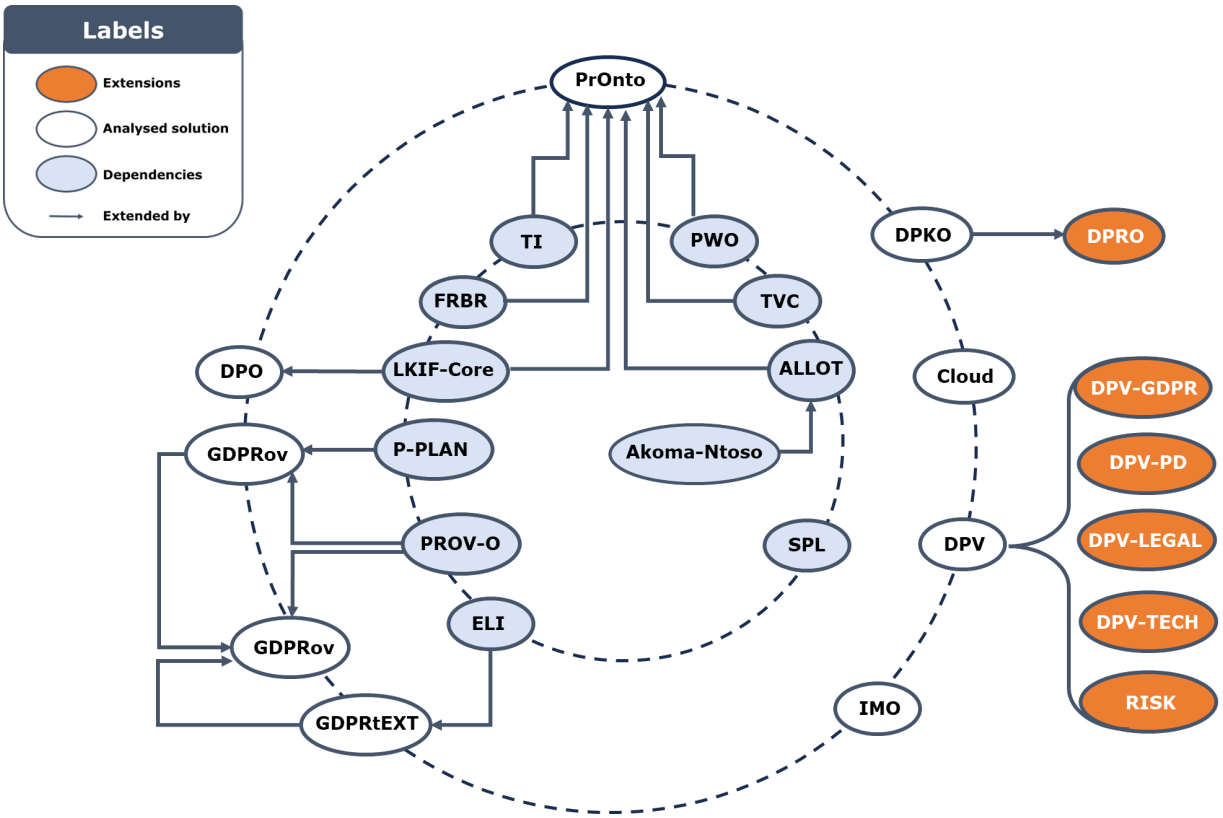


Figure 2.2: Data protection vocabularies dependency chart.

least one principle. For instance, data subjects have the right to rectify inaccurate or incomplete data, and data controllers must provide the means to do it, according to GDPR's *'accuracy'* principle. Furthermore, DPO defines consent as a legal justification connected with the principle of trust and also specifies concepts to model the special case of giving consent as a parent for a child, although the concept of *'consent provided by delegation'* is missing. Data protection-related entities, such as data controllers, supervisory authorities, processors, representatives, or data protection officers, are modelled as a type of **person**.

Listing 2.6 provides an example of a SPARQL query to retrieve duties of a data controller that do not concern the transfer of personal data using DPO.[23]

GDPRov

The GDPR Provenance Ontology (GDPRov) [Pandit and Lewis, 2017] aims to record the provenance of personal data and of the consent conditions and processing activities performed over such data, according to the GDPR. GDPRov extends PROV-O [Lebo et al., 2013], a W3C Recommendation created to define the provenance of entities and systems, and P-Plan [Garijo and Gil, 2012], an extension of PROV-O to represent activities and corresponding steps to execute them, as well as the entities involved. Using these terms, it is possible to monitor changes in consent or to track

[23] More examples are available in a project repository at `https://bitbucket.org/guerret/lu.uni.eclipse.bpmn2/` (accessed on 16/July/2023).

Listing 2.6 SPARQL query to retrieve duties of the data controller that do not concern data transfer using the Data Protection Ontology.

```
SELECT DISTINCT ?duty WHERE {
    ?x rdfs:subClassOf* [ rdf:type dpo:Rule ] .
        ?x rdfs:label ?duty
        MINUS {
                ?x rdfs:subClassOf* [ rdf:type dpo:TransferRule ] .
                ?x rdfs:label ?duty .
        } .
        FILTER (?duty != "Rule"@en && ?duty != "ProcessingRule"@en &&
         ↪  ?duty != "TransferRule"@en) }
```

Listing 2.7 SPARQL query retrieving entities involved in acquiring consent using GDPRov [Pandit and Lewis, 2017].

```
SELECT ?consent ?template ?toc WHERE {
    ?consent a gdprov:ConsentAgreement .
    ?template a gdprov:ConsentAgreementTemplate .
    ?toc a gdprov:TermsAndConditions .
    ?step a gdprov:ConsentAcquisitionStep .
    ?step gdprov:usesConsentAgreementTemplate ?template .
    ?step gdprov:usesTermsAndConditions ?toc .
    ?step gdprov:generatesConsentAgreement ?consent }
```

the interaction between entities involved in the exchange of data.

GDPRov's main concept to record the provenance of consent is the **consent agreement template** class, a common template that includes the consent conditions presented to the users and the entities in charge of data processing, and also information on third party sharing, approved processing activities and additional rights. Provenance metadata on the origin, use, storage, and sharing of the data can also be recorded with GDPRov, as well as information regarding transformations performed to the data. In addition to this, provenance data on the exercising and fulfilment of GDPR-related rights and obligations can also be represented with GDPRov – for each right or obligation, a plan can be modelled to include the activities that need to be executed when a user exercises a particular right.

Listing 2.7 illustrates a SPARQL query that uses GDPRov's terms to retrieve the entities involved in acquiring consent.

Cloud

The Cloud GDPR ontology was developed by Elluri and Joshi [2018] to express data protection obligations of cloud data consumers and providers, taking into consideration the Cloud Security Alliance (CSA) controls defined on the *Code of Conduct for GDPR Compliance* [Privacy Level Agreement Working Group, 2017].

The **stakeholders**, **controls**, and **obligations** are the core concepts of this ontology. Cloud-related obligations are extracted and connected to GDPR's articles and are also associated with

CSA requirements using the **hasCSAcontrol** property. Moreover, these obligations are further specified into common or provider/consumer-specific obligations taking into consideration which stakeholders they are applicable to, e.g., maintaining records of data processing activities and notifying data breaches are common obligations, while providing legal representatives for non-EU stakeholders or hiring a DPO are the responsibility of the consumer and of the provider, respectively.

This work was later extended [Elluri et al., 2018] to automate the implementation of compliance tasks mandated by the GDPR and the Payment Card Industry Data Security Standard (PCI DSS) guidelines [PCI Security Standards Council, 2018] and also to include the rights of consumers, providers and end users. These guidelines deal with financial data, such as the credit card number or card-holder's name, which fall under the protection offered by the GDPR. As such, and since its scope is narrower, a data breach of PCI DSS-related data automatically results in a GDPR-related one. Thus, the cloud-related PCI DSS requirements were used to enhance this ontology and its evaluation was performed using privacy policies from five major companies that deal with card-holder's data.

PrOnto

PrOnto, the Privacy Ontology for legal reasoning, aims to model GDPR-related associations between agents, processing activities, data categories, and deontic modalities, with the main goal to support legal reasoning and compliance with data protection regulations [Palmirani et al., 2018]. PrOnto relies on a number of ontologies to model these relationships:

- LKIF-Core [Hoekstra et al., 2007] is used to model agents, e.g., organisations, software, or people, as well as the several legal roles which can be assigned to them, i.e., acting as a data controller or processor.
- The Functional Requirements for Bibliographic Records (FRBR) ontology [Byrum et al., 2009] is used to model legal documents as sources of information, that regulate the different relationships between the agents documented in the text, and to register changes in their representation over time.
- The ALLOT (A Light Legal Ontology On Top level classes) ontology, developed by [Barabucci et al., 2010], extends the Akoma Ntoso standard [Palmirani and Vitali, 2011] to link documents with the data they contain, e.g., people, events or locations.
- The Publishing Workflow Ontology (PWO) [Gangemi et al., 2017] is used to characterise provenance data associated with the publication of a document, e.g., to model the different types of processing of personal data.
- The TVC (Time-indexed Value in Context) ontology [Peroni, 2014] and the TI (Time Interval) ontology pattern[24] are used to associate time-dependent variables such as events to specific agent roles that only emerge in particular point in time.

PrOnto's main classes are defined as follows: **documents and data**, **agents and roles**, **processing**

[24] http://www.ontologydesignpatterns.org/cp/owl/timeinterval.owl (accessed on 16/July/2023)

and workflow, **legal rules and deontic formula**, and **purposes and legal bases**. GDPR is the main document used to extract personal data categories such as judicial or sensitive data and non-personal data such as anonymous or legal person data. Moreover, processing activities are represented as a workflow of actions that should be recorded together with provenance data such as the context and time in which each action occurs. Each processing activity is also associated with a purpose and a legal rule, which is composed of deontic specifications, i.e., prohibitions, rights, permissions, and obligations, used to check if the activity being executed is compliant with the GDPR. PrOnto is, however, not publicly available for re-usage, which hinders the assessment of the totality of concepts it models[25].

Listing 2.8 illustrates a query to retrieve information regarding personal data processing activities performed by company *X* in the role of data controller from time *t1* to *t2*.

Listing 2.8 SPARQL query used to retrieve personal data processing activities performed by company *X* in the role of data controller from *t1* to *t2* using PrOnto [Palmirani et al., 2018].

```
SELECT ?pdp WHERE {
    ?pdp pronto:isManagedBy _:c .
    [ lkif:plays _:c ;
        rdfs:label "X" ] .
    ?pdp pronto:isValid [
        ti:hasIntervalStartDate [
            a ti:TimeInterval; rdfs:label "t1" ] ;
        ti:hasIntervalEndDate [
            a ti:TimeInterval; rdfs:label "t2" ] ] . }
```

GConsent

GConsent [Pandit et al., 2019a] is an ontology focused on the GDPR concept of consent and, as such, it models information regarding how consent was collected and stored, as well as records of any changes that may occur over time, including withdrawal, including data on the parties involved, according to Article 6 of the GDPR. Additionally, other documentation sources were also adopted, including EDPB's guidelines on consent [European Data Protection Board, 2020a]. GConsent encompasses not just the notion of consent, but also depicts its status, context, and origin. To achieve this, it leverages established vocabularies in this domain, including PROV-O [Lebo et al., 2013], GDPRov [Pandit and Lewis, 2017], and GDPRtEXT [Pandit et al., 2018] (discussed in Section 2.2.2).

GConsent's main concepts are the terms to represent **data subjects**, **personal data**, **purpose** and **processing types**, as well as the **consent** and **consent status** classes – the latter fills a gap on the available ontologies on this domain as it not only defines **explicitly given consent** as a concept, but also classifies other states of consent, such as **implicitly given**, **expired** or **refused**. GConsent also includes concepts to represent information regarding the context in which the consent was obtained – information about the entities involved, spatial and temporal aspects, and

[25] As PrOnto is not available online, its namespace is unknown. In this Thesis, the prefix `pronto` is used to identify PrOnto's concepts that are available in the analysed publication.

the format used to record it, e.g., Web form, voice recording, or signature – and a taxonomy of processing types, e.g., data alignment or data retrieval.

Listing 2.9 provides an example of a record of provenance data regarding an activity of withdrawing consent using GConsent.

Listing 2.9 Turtle record of provenance data regarding an activity of withdrawing consent using GConsent [Pandit et al., 2019a].

```
ex:modifyConsent a gdprov:ModifyConsentActivity ;
    prov:invalidated ex:consent1 ;
    prov:generated ex:consent2 .

ex:consent1 a gc:Consent ;
    gc:isPreviousConsentFor ex:consent2 ;
    gc:hasStatus gc:ConsentStatusExplicitlyGiven .

ex:consent2 a gc:Consent ;
    gc:hasStatus gc:ConsentStatusWithdrawn .
```

IMO

The Information Model Ontology (IMO) [Papagiannakopoulou et al., 2014, Lioudakis and Cascone, 2019] was developed in the context of the BPR4GDPR project[26] and it aims to define the entities and respective roles that are involved in the organisation lifecycle of processing personal data.

IMO's main concepts encompass various categories: **data types**, **roles** across diverse organisational structures, **machine types** housing **operations** tailored for specific **purposes**, **events** alongside their contextual details. The roles class specifically pertains to the duties assigned to users within organisational contexts, with the potential for hierarchical implementations based on the granularity of data. Data processing activities are represented via the operations class, which is equipped with the **hasInput** and **hasOutput** properties. These properties facilitate the connection between operations and the data being processed, as well as the resulting data and their corresponding states, e.g., normal or anonymised form. These operations can be organised within an operation container, a concept designed to group processing activities that typically function together within specific contexts. For example, within database management, functions such as create, read, update, or delete are frequently employed together. Instances of the role and operation classes are consistently linked with a purpose instance. Moreover, the events class, designed to encompass all processing activities such as data breaches or consent withdrawals, should be paired with the context class to create specific event instances with temporal and spatial details, among other pertinent information.

[26] BPR4GDPR (Business Process Re-engineering and functional toolkit for GDPR compliance) is a European Union H2020 innovation programme with the main goal of providing a framework to reinforce the implementation of GDPR-compliant measures inside organisations at diverse scales and in several domains. More information is available at https://www.bpr4gdpr.eu/ (accessed on 16/July/2023).

DPV

The Data Privacy Vocabulary is being developed and maintained by the W3C Data Protection Vocabularies and Controls Community Group (DPVCG)[27] – an output of a W3C workshop on data privacy controls which took place in 2018, with the objective of defining priorities for the standardisation of this domain [Bonatti et al., 2018b].

DPV's core concepts, **processing**, **purpose**, **recipient** and **personal data categories**, were adapted from the SPECIAL Usage Policy Language (SPL) vocabularies [Bonatti et al., 2018a] and new concepts were/are continuously added to DPV after being discussed and agreed upon by the Community Group. In addition to previously mentioned concepts, DPV's base vocabulary was first published with the following additional classes: **legal basis**, **technical and organisational measures** and **legal entities**, including **data subject** and **child**, **recipients**, **data controller**, **data processor** and **third party** [Pandit et al., 2019b]. In December of 2022, DPV version 1 was published including additional concepts to model **risk**, **rights** and **data subject rights**, **technology**, **laws** and **context of processing** classes and the previously existing taxonomies were extended with new terms – new legal entities, including authority and data protection authority, vulnerable data subject, data sub-processor, data protection officer and representative, were added to the vocabulary, as well as new purpose and legal basis sub-classes. For instance, the purpose taxonomy is composed of 76 purpose sub-classes, which are topped by classes such as R&D or Service Provision, and can be further constrained to specific contexts or business sectors, and DPV's processing operations taxonomy covers the terms defined in Article 4.2 of the GDPR, providing a set of 44 processing categories.

DPV also provides five extensions:

- Personal data categories are defined in the PD extension[28] – the personal data class is split into generic classes such as financial or social data, adapted from the **EnterPrivacy** taxonomy by Cronk [2017], which are further specified into concepts such as credit card number or social media accounts.
- GDPR-specific concepts are defined in the EU-GDPR extension[29] – it covers legal bases specified on GDPR's Articles 6 and 9 for the processing of personal data and also the legal bases for the transfer of personal data to third countries defined on Articles 45, 46 and 49. It also models GDPR's data subject rights, data transfer tools, and Data Protection Impact Assessment (DPIA) terms.
- Technology-relevant concepts are defined in the TECH extension[30] – it includes concepts to model specific technologies, their management, technology actors, and relevant tools and systems.
- Jurisdiction-relevant concepts are defined in the LEGAL extension[31] – it contains terms related to specific laws, adequacy decisions, and a taxonomy of authorities.

[27] https://www.w3.org/community/dpvcg/ (accessed on 16/July/2023)
[28] https://w3id.org/dpv/pd (accessed on 17/July/2023)
[29] https://w3id.org/dpv/legal/eu/gdpr (accessed on 17/July/2023)
[30] https://w3id.org/dpv/tech (accessed on 17/July/2023)
[31] https://w3id.org/dpv/legal (accessed on 17/July/2023)

- Risk concepts are defined in the RISK extension[32] – it covers concepts related to risk, likelihood, severity levels, consequences, and impacts, as well as risk assessment techniques and methodologies.

Listing 2.10 provides an example of a personal data handling related to the collection and usage of email addresses for marketing purposes using DPV.

Listing 2.10 Turtle record of a personal data handling related to the collection and usage of email addresses for marketing purposes using DPV [Pandit et al., 2019b].

```
ex:PDHforMarketing a dpv:PersonalDataHandling ;
    dpv:hasDataController [ a dpv:DataController ] ;
    dpv:hasPersonalData pd:EmailAddress ;
    dpv:hasProcessing dpv:Collect, dpv:Use ;
    dpv:hasPurpose dpv:Marketing .
```

GDPRtEXT

GDPRtEXT was developed by Pandit et al. [2018] as a linked data resource, which extends the European Legislation Identifier (ELI) ontology [Office of Publications on Eur-Lex, 2017], to connect GDPR concepts with the specific chapters, articles or points of the regulatory text.

The main concepts modelled in this ontology are related to the specific **entities** mentioned in the GDPR text, their **rights** and **obligations**, the **principles** and the **activities** which specify processes and actions defined in the GDPR, such as reporting a data breach, exercising rights or demonstrating consent. These terms are linked to the relevant GDPR provisions using `rdfs:isDefinedBy`.

2.2.3 Comparative analysis

As it is visible in Tables 2.4 and 2.5, existing ontologies and vocabularies in the domain of data protection are of particular interest to represent the privacy terms described in Table 2.2, however, there are still representational gaps to be filled. More specifically, each ontology was analysed in terms of the concepts they model and, when a particular concept was found to represent a particular privacy term, the name of the respective term was included in the comparative tables, as well as the number of sub-classes which can be used to more precisely define the term. The cases where no concepts were found to represent the privacy term, but existing terms could be expanded to do it, are indicated with an asterisk. Privacy terms I15, I19, I24, I26, and I28 to I30 are absent from both Table 2.4 and 2.5 since none of the analysed ontologies include them. This can be justified by the lack of vocabularies focusing on contractual-based personal data processing activities, in the case of I15 which relates to contractual obligation details, and on rights exercising activities, in the case of I19, I26, I29, and I30 which are justifications for data controllers to not comply with data subject rights. Moreover, I24 and I28 are specific terms, related to technical measures to erasure data and transfer data between controllers, respectively, which seem to lack representation in machine-readable vocabularies as there are no specific solutions focusing on data erasure or portability, with most vocabularies focusing on the terms included in the *'right to be informed'* (Arts. 13 and 14) [2016b].

[32] `https://w3id.org/dpv/risk` (accessed on 17/July/2023)

Table 2.4: Representation of privacy terms I1 to I22 in the DPKO, DPRO, DPO, GDPRov, Cloud, and PrOnto ontologies. The names of the classes that can be used to specify a particular item are depicted in the table, as well as their respective number of subclasses, if available. The privacy terms that can not be fully represented by the current ontology terms are illustrated with an asterisk.

	DPKO	DPRO	DPO	GDPRov	Cloud	PrOnto
I1			Controller	Controller		*
I3				ControllerRepresentative	Identify_Representatives	
I6	Purpose		Purpose			Purpose (10)
I7	*		LegalJustification (6)			
I8			LegitimateInterest			
I9			Recipient (2)			*
I10				*		
I11						*
I12			DataSubjectRight (7)	Process (10)		Right (8)
I16			AutomatedProcessing	*		
I17	*		PersonalData	PersonalData (3)		PersonalData (7)
I20			*			
I21				ProvideCopyOfPersonalData		
I22				RectifyData		

DPO, GDPRov, PrOnto, DPV, and GDPRtEXT offer the potential to partially populate a significant portion of the informational items necessary for the *'right to be informed'* (Arts. 13 and 14) [2016b] and other data subject rights (Arts. 15 to 22) [2016b]. It is important to emphasise DPV and GDPRtEXT as they include concepts to represent, at least partially, 19 and 14 informational items, respectively, out of the 30 outlined in Table 2.2. Moreover, these vocabularies boast the highest count of subclasses dedicated to specifically defining their respective items.

It should be noted that the majority of the resources presented are outdated or lacking recent developments, with DPV being the sole exception, having introduced new outcomes in the past two years. Additionally, among all the covered vocabularies, only DPKO, IMO, and PrOnto lack open and accessible resources.

2.3 Using policy languages to specify access conditions

Policy languages have been used in the last decades to specify information regarding the usage of data, e.g., to represent licenses associated with datasets or software usage. As such, they seem perfectly aligned with the goal of representing the conditions to access data on the Web, whether being preferences set by data subjects or access requests from other entities. In addition, if used together with privacy and data protection-specific terms, e.g., coming from the ontologies described in the previous Section, they can be used to model legally-aligned access control policies.

In the following Sections, the criteria used to analyse each specified policy language are described, as well as a description of each identified language.

2.3.1 Criteria for analysis

Each solution is accompanied by an introductory summary of the language, detailing its primary contributions, followed by an overview of its core elements. Additionally, where applicable, specific

Table 2.5: Representation of privacy terms I1 to I27 in the GConsent, IMO, DPV, and GDPRtEXT ontologies. The names of the classes that can be used to specify a particular item are depicted in the table, as well as their respective number of subclasses, if available. The privacy terms that can not be fully represented by the current ontology terms are illustrated with an asterisk.

	GConsent	IMO	DPV	GDPRtEXT
I1	DataController	DataController	DataController	Controller
I2			hasContact	
I3			Representative	ControllerRepresentative
I4			hasContact	
I5			hasContact	
I6	Purpose	Purposes	Purpose (76)	
I7	*		LegalBasis (26)	LawfulBasisForProcessing (14)
I8			A6-1-f	LegitimateInterest
I9	*		Recipient (4)	*
I10			*	CrossBorderTransfer
I11	*	*	*	RecordDataRetentionPeriod
I12			DataSubjectRight (12)	Rights (10)
I13	*		A7-3	
I14			A77	
I16			AutomatedDecisionMaking	AutomatedProcessing
I17		DataTypes (52)	PersonalData (206)	PersonalData (5)
I18			DataSource	InfoAboutSourceOfData
I20			*	
I21				ProvideCopyOfPersonalData
I23				RightOfErasure (2)
I25			hasContact	
I27				RightToRestrictProcessing (3)

examples of use cases employing the language are provided, along with details on any derived implementations, including information on available reasoners that use it. The dependencies on prior existing works are also noted when outlined in the literature. Table 2.6 presents a concise description of the policy languages detailed in the following Sections, along with details about the creators of the resources, version number, publication date, and the most recent update date. The solutions were examined chronologically based on their publication date, followed by their last update date. Figure 2.3 depicts a dependency graph illustrating the relationships between languages, their dependencies, and subsequent developments.

Moreover, the following criteria were used to analyse existing research on semantic policy languages:

(C1) Ability to model deontic concepts, e.g., permissions, prohibitions, obligations.

(C2) Ability to model GDPR concepts, such as the privacy terms in Table 2.2.

(C3) Existence of taxonomies of terms to populate policy conditions.

(C4) Existence of mechanisms to assist with compliance.

Table 2.6: Brief description of the policy languages described in Section 2.3.2.

Abbreviation (Section)	Full Name	Creators	Version	Date of publication	Last update
P3P (2.3.2)	Platform for Privacy Preferences	Cranor et al.	1.0	1998	2010
ODRL (2.3.2)	Open Digital Rights Language	Iannella et al.	2.2	2001	2019
XPref (2.3.2)	XPath-based Preference Language	Agrawal et al.	-	2003	-
AIR (2.3.2)	Accountability In RDF	Khandelwal et al.	-	2007	2009
S4P (2.3.2)	SecPAL for Privacy	Becker et al.	-	2009	2010
POL (2.3.2)	Privacy Option Language	Stefan Berthold	-	2010	2013
PPO (2.3.2)	Privacy Preference Ontology	Sacco and Passant	-	2011	2013
LegalRuleML (2.3.2)	LegalRuleML Core Specification	Palmirani et al.	1.0	2012	2021
A-PPL (2.3.2)	Accountable Policy Language	Azraoui et al.	-	2013	2016
P2U (2.3.2)	Purpose-To-Use	Iyilade and Vassileva	-	2014	-
SPL (2.3.2)	SPECIAL Usage Policy Language	Bonatti et al.	1.0	2017	2019
DPF (2.3.2)	Declarative Policy Framework	Martiny et al.	-	2018	2020
LPL (2.3.2)	Layered Privacy Language	Gerl et al.	-	2018	2019

(C5) Resource is maintained/continues to be actively developed.

(C6) Existence of an open and accessible specification.

The outcomes of this comparative analysis will be provided in Section 2.3.3 and systematised in Table 2.7.

2.3.2 Semantic policy languages for access control

This Section's main goal is to describe existing policy languages related to privacy, delineating the structure and information offered by each language. Additionally, their compatibility with the GDPR is assessed, focusing on their ability to describe the provisioned rights and obligations. Moreover, Table 2.6 provides an overview of these languages and collects information about the creators of the resources, their versions, the date of publication, and the date of the last known update. Said languages are then analysed in chronological order regarding the date of publication and a dependency graph is presented in Figure 2.3.

To complement the description of the languages presented in this Section, additional documentation and resources were published on a Web page[33], including diagrams and code examples.

P3P

Cranor et al. [2002b] introduced the Platform for Privacy Preferences (P3P) language as a standard for Web services to disclose their privacy practices in a machine-readable format. This facilitated user agents in easily interpreting these practices and notifying users about decisions based on them. However, despite enabling users to be informed about the privacy policies of Web pages, these mechanisms do not ensure that the pages are actively adhering to these policies, as P3P lacks enforcement capabilities. Therefore, the P3P vocabulary was designed not to comply with a

[33]Available at https://w3id.org/people/besteves/phd/sota/languages. Its public repository can be consulted at https://w3id.org/people/besteves/phd/sota/repo for further improvement when new solutions appear.

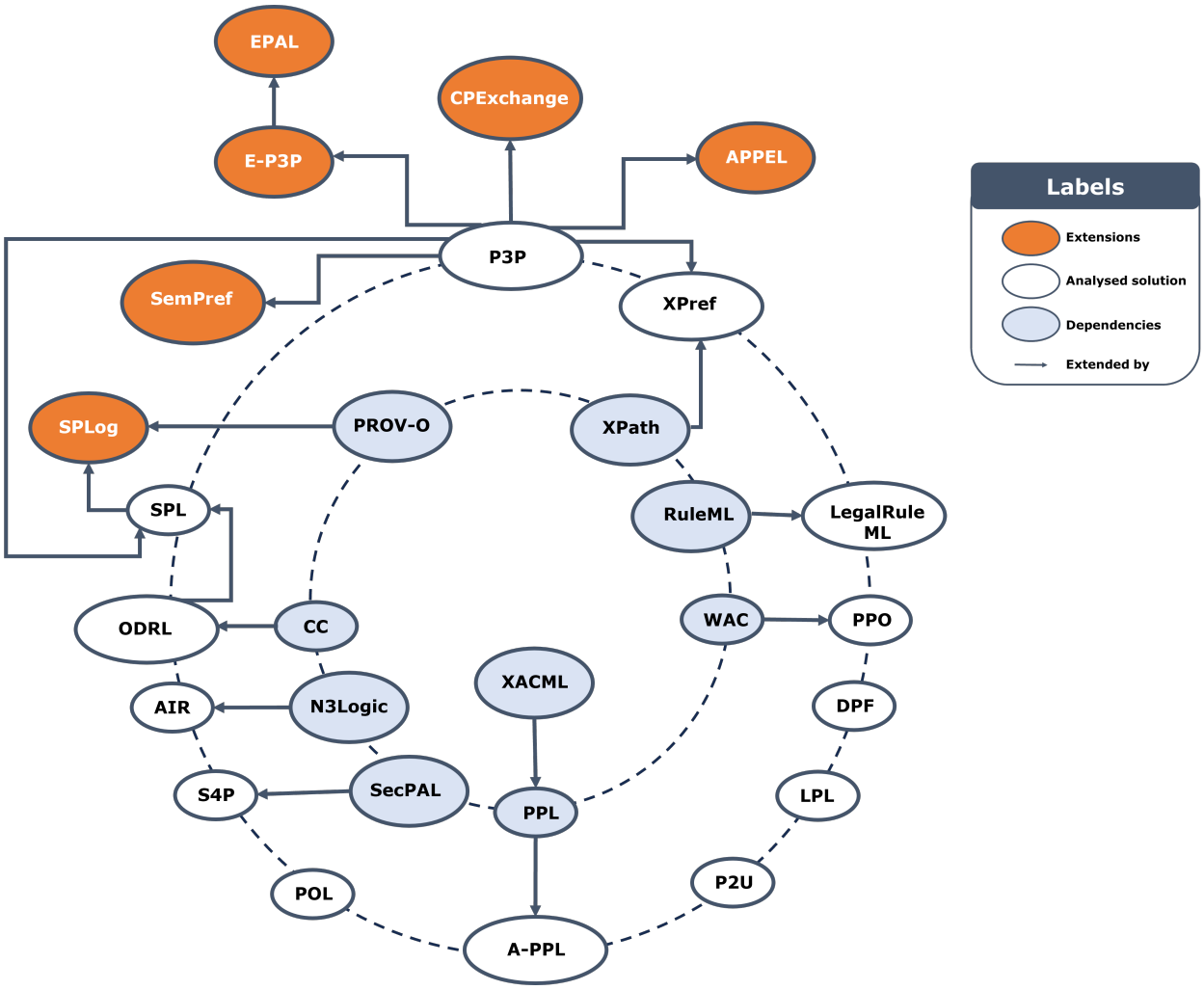


Figure 2.3: Privacy-related policy languages dependency chart.

specific regulation but rather to specify the practices of single Web pages.

The primary contributions of the P3P specification include a data schema for outlining the data intended to be collected by the Web page, a standardised set of purposes, data categories, and recipients, and an XML standard for defining privacy policies. P3P policies consist of both general assertions and specific statements, the latter being related only to certain types of data. General assertions encompass the legal entity applying the policy and informational elements related to access, disputes, and remedies. The access element indicates whether the Web page allows access to the data it gathers and the disputes element outlines a process for resolving privacy-related disputes, while the remedy element details potential solutions in the event of a policy breach. Additionally, each P3P statement consists of a distinct 'data group' containing one or more data elements, along with purpose, recipient types, and retention elements. In this context, P3P outlines a range of purposes relevant to Web-based data processing operations, such as facilitating and supporting the initial activity for which the data was supplied, conducting research and development, or performing data analysis. The recipient type element can be used to specify who will benefit from the collected data, while the retention element must accurately reflect the policy regarding how

Listing 2.11 P3P policy extracted from Example 3.1 of the P3P specification [Cranor et al., 2002b], which specifies the privacy policy of CatalogExample.

```
<http://example.com/#forBrowsers> a p3p:Policy ;
    p3p:disclosure <http://example.com/PrivacyPractice.html> ;
    p3p:entity [
        p3p:business.name [ rdf:value "CatalogExample" ] ;
        p3p:business.contact-info.postal.street [
            rdf:value "4000 Lincoln Ave." ] ;
        p3p:business.contact-info.postal.city [
            rdf:value "Birmingham" ] ;
        p3p:business.contact-info.postal.stateprov [ rdf:value "MI" ] ;
        p3p:business.contact-info.postal.country [ rdf:value "USA" ] ;
        p3p:contact.online.email [ rdf:value "catalog@example.com" ] ;
        p3p:contact.telephonenum.intcode [ rdf:value "1" ] ;
        p3p:contact.telephonenum.loccode [ rdf:value "248" ] ;
        p3p:contact.telephonnum.number [ rdf:value "3926753" ] ] ;
    p3p:access p3p:AccessClass-nonident ;
    p3p:statement [
        p3p:purposeAlways p3p:Purpose-admin, p3p:Purpose-develop ;
        p3p:recipientAlways p3p:Recipient-ours ;
        p3p:retention p3p:Retention-stated-purpose ;
        p3p:data [
            rdf:predicate p3p:dynamic.clickstream, p3p:dynamic.http ] ] .
```

long the data will be kept. Listing 2.11 displays a P3P policy that underscores the aforementioned P3P elements. CatalogExample gathers essential details concerning its users' computer systems, as well as data regarding the pages they visit. This information serves system administration and research and development purposes and is exclusively utilised by the company and retained for a duration deemed suitable for the specified purposes.

P3P was initially developed to articulate policies of Web services, prompting the design of APPEL as an extension to empower users in expressing their preferences [Cranor et al., 2002a]. Consequently, the utilisation of both languages becomes imperative to align user privacy preferences with service privacy policies. Furthermore, Bohrer and Holland [2000] introduced the CPExchange language, an XML specification facilitating the transfer of customer data across enterprise services, incorporating P3P privacy policies relevant to the exchanged data. Likewise, EPAL [Ashley et al., 2003], developed by IBM Research[34] along with its precursor E-P3P [Ashley et al., 2002], leveraged P3P statements to align enterprise privacy policies with user preferences. Li et al. [2006] introduces a declarative data-centric semantic model alongside a succinct syntax for P3P policies, facilitating the representation of the relationship between various P3P elements. The primary aim of this language is to articulate policies in a manner that can be uniformly interpreted and represented across diverse user agents. Extending this semantic foundation, the authors put forward SemPref, a preference language that considers the significance of the privacy policy rather than its syntactic form.

The P3P 1.0 Specification achieved W3C Recommendation status on April 16, 2002. Nonetheless, its

[34] http://www.research.ibm.com/ (accessed on 16/July/2023)

adoption was restricted as it requires acceptance from both Web services and users. Furthermore, there has been no protocol established for these P3P policies to accurately reflect the privacy practices of Web pages. While this specification did attain W3C recommendation status, its failure to gain widespread adoption rendered it obsolete by 2018. Nonetheless, the significance of P3P remains considerable, as its inception and utilisation marked a pioneering endeavour in the realm of machine-readable privacy languages. Consequently, the primary lessons derived from this language pertain to the necessity of establishing a formal semantics capable of delineating both data subject and controller policies, which accurately reflect their data preferences and practices, and the need for tools that effectively enforce the outlined policies.

ODRL

The ODRL Vocabulary & Expression 2.2 [Iannella and Villata, 2018] gained W3C Recommendation status in February 2018, developed by the Permissions & Obligations Expression Working Group, with its initial version launched back in 2001. Its primary objective was to establish a language capable of translating natural language policies into machine-readable formats, specifying details regarding permissions, prohibitions, and obligations pertaining to an asset. This vocabulary stems from the consolidation of previous efforts undertaken by the ODRL CG, encompassing the ODRL V2.1 Common Vocabulary, the ODRL V2.1 XML Encoding, the ODRL V2.1 Ontology, and the ODRL V2.1 JSON Encoding. Ongoing maintenance of the ODRL's standards and specifications are supported by the ODRL CG.

ODRL includes two vocabularies for the description of policies: the ODRL Core Vocabulary and the ODRL Common Vocabulary. The primary class within ODRL's Core Vocabulary is the 'policy' concept, facilitating the identification of a specific policy through its unique identifier. Within each policy, there may exist multiple rules – an abstract class that outlines the shared characteristics of permissions, prohibitions, and duties. These rule types serve to declare whether a particular action, e.g., over as asset, is permitted, prohibited, or obligatory. Additionally, permissions might also be linked with duties that must be fulfilled for said permissions to be active. Furthermore, rules undergo further refinement through the usage of constraints, which specify the circumstances under which the rule applies, e.g., a particular permission remains valid until the conclusion of 2024. The ODRL Vocabulary also outlines a collection of 49 actions, nine of which are imported the Creative Commons (CC) vocabulary. The entities, or parties, involved (which can encompass a group of individuals, an organisation, or an agent) are responsible for enforcing the rules and may assume various roles contingent upon their relationship with the asset, e.g., the entity issuing the rule adopts the assigner role, whereas the recipient of the rule assumes the assignee role. An asset, on the other hand, refers to an identifiable resource, such as data, software, services, or a combination thereof, that is subject to a rule. The ODRL Common Vocabulary further delineates subclasses of policies, roles played by the involved parties, taxonomies of use and transfer actions for rules, and a variety of constraint operands, such as temporal, spatial, or sector-specific. Of particular significance concerning the GDPR is the privacy policy subclass regarding assets containing personal data. Consequently, privacy policies implementing the ODRL language must specify to the involved parties the manner in which data is utilised, as well as with whom and for what purpose. Listing 2.12 provides an implementation of an ODRL privacy policy that underscores the aforementioned elements, including a duty for the assignee to be allowed to use the data and a consequence in case they do not.

Listing 2.12 ODRL `Privacy` policy.

```
<http://example.com/privacy-policy> a odrl:Privacy ;
    odrl:uid <http://example.com/privacy-policy> ;
    odrl:permission [
        odrl:target <http://example.com/beatriz/contacts> ;
        odrl:assignee <http://example.com/beatriz> ;
        odrl:assigner <http://example.com/company-a> ;
        odrl:action odrl:use ;
        odrl:duty [
            odrl:action odrl:obtainConsent ;
            odrl:consentingParty <http://example.com/beatriz> ;
            odrl:consequence [
                odrl:assigner <http://example.com/company-a> ;
                odrl:action odrl:delete ] ] ] .
```

ODRL's representational capabilities exhibit some limitations, as highlighted by Kebede et al. [2021], particularly regarding the portrayal of delegation, the various semantics utilised for expressing duties, and the management of conflicts. Nevertheless, efforts such as those described by Fornara and Colombetti [2018, 2019] are underway to formalise the semantics of ODRL policies.

ODRL was applied in various contexts, including its usage by working groups within the Open Mobile Alliance SpecWorks[35] and by the IPTC Rights Expressions WG for the RightsML Standard[36].

XPref

Agrawal et al. [2005-08] introduced XPref as an alternative to APPEL, which only permits the definition of P3P policies that are not allowed by the user. XPref utilises XPath (XML Path Language) 1.0 and 2.0 expressions to replace APPEL rules, enhancing the precision and reducing errors in policy formulation. Both XPath 1.0, as described by Clark and DeRose [1999], and XPath 2.0, as detailed by Berglund et al. [2010], attained W3C Recommendation status on November 16th, 1999, and December 14th, 2010, respectively. It should be noted that these specifications are no longer subject to further maintenance since subsequent versions have been developed. In this context, the primary objective of XPath is to offer a method for traversing the hierarchical elements within an XML document. In pursuit of this objective, XPath views an XML document as a tree structure of nodes When an XPath expression is applied to the document, it determines an ordered sequence of nodes, resulting in a concise path representation. This path consists of expressions that yield various types of nodes, including root, element, text, attribute, namespace, processing instruction, or comment nodes.

XPref was crafted to ensure that its rules not only recognise combinations of P3P elements that render a policy unacceptable based on user preferences, but also confirm that the presented elements are defined as acceptable. It achieves these objectives while preserving the syntax and semantics of APPEL, along with its core classes. However, the contents of the rules are substituted with XPath expressions, given that P3P policies are XML documents and can thus be readily

[35] https://www.omaspecworks.org/ (accessed on 18/July/2023)

[36] https://www.iptc.org/std/RightsML/2.0/RightsML_2.0-specification.html (accessed on 18/July/2023)

compared with XPath-based rules. These expressions are defined by appending a 'condition' attribute to the rule, which activates the rule when the XPath expression yields a non-empty result.

AIR

In 2010, Khandelwal et al. developed Accountability in RDF (AIR) – a declarative language enabling the assertion of facts and the inclusion of rules. AIR is built on N3Logic [Berners-Lee et al., 2008], which supports rule nesting, rule reuse, and automated explanations of actions carried out by the AIR reasoner. These explanations can be customised and, given that they may contain sensitive data like Personally Identifiable Information (PII), can be employed to ensure privacy. For example, they can be utilised to conceal actions executed under specific rules.

N3Logic extends the RDF data model with the aim of expressing logic rules on the Web, thereby promoting the use of a unified language for both data representation and logical inference. Thus, AIR leverages N3Logic's inherent capabilities, including built-in functions, nested graphs, and contextualised reasoning. This enables AIR rules to incorporate the usage of graphs as literal values, and built-in functions or operators defined as RDF properties.

Each AIR rule is assigned a unique IRI, ensuring its seamless integration with the linked data cloud and facilitating its reuse. These rules adhere to the following structure: `air:if condition`; `air:then then-actions`; `air:else else-actions`. The action instances can include annotations using the `air:description` property. These annotations are subsequently integrated by the AIR reasoner into its justifications and can serve to conceal PII found within the rules. Additionally, the format of the rules graph permits the nesting of rules within the same rule set. This feature offers a means to segment the conditions outlined by the rule, allowing only a portion of them to be revealed in the justifications.

S4P

S4P (SecPAL for Privacy), designed by Becker et al. [2009, 2010], constitutes a language framework designed for articulating users' privacy preferences and the data handling practices of Web services. Originating from Microsoft Research[37], this language serves as an extension of the company's earlier endeavor, SecPAL, aimed at delineating PII management.

SecPAL [Becker et al., 2007] is a flexible, decentralised authorisation language, crafted for articulating policies and enhancing their expressiveness to define delegation conditions, domain-specific constraints, and negation. An authorisation policy comprises a set of assertions, each associated with an issuer responsible for vouching for the assertion, along with a collection of conditional facts and constraints pertaining to temporal or spatial aspects. Subsequently, when an access request is made, it undergoes a transformation into a series of queries. These queries are then matched against the clauses that represent the system's policies, ultimately leading to a decision on data access conditions. S4P extends SecPAL by treating granted rights and required obligations as assertions and queries. Based on these, a satisfaction checking algorithm is formulated to evaluate the disclosure of PII between users and data-collecting services. As a result, services should articulate their data-handling practices in the form of SecPAL queries. Conversely, users

[37] `https://www.microsoft.com/en-us/research/` (accessed on 16 March 2024)

Listing 2.13 S4P example, extracted from Becker et al. [2009], which specifies Alice's privacy preferences concerning the collection of her email address by eBooking services.

```
Alice says x may use Email for p if
    x is a eBookingService,
    where p ∈ {Confirmation, Newsletter, Stats}
Alice says x may send Email to y if
    x is a eBookingService,
    y is a TrustedPartner
Alice says x can say y is a TrustedPartner if
    x is a eBookingService
Alice says ⟨Service⟩ is a RegisteredService? ∧
    ∃t (⟨Service⟩ says ⟨Service⟩ will delete Email within t? ∧ t ≤ 30 days?)
```

specify their preferences as SecPAL assertions, precisely delineating what services are authorised to do and what duties they have regarding r=the usage of said PII. If the algorithm yields a positive result, indicating that the service's policies complies with the user's preferences, the service can proceed with its data handling activities. Additionally, S4P establishes a data disclosure protocol to ensure that users' preferences are respected when their data is shared with third party recipients.

In addition to possessing an XML schema for implementation purposes, S4P features a human-readable and unambiguous syntax, enabling its utilisation in various applications. Listing 2.13 illustrates the S4P syntax in a scenario where Alice, the user, delineates her privacy preferences concerning the collection of her email address. Specifically, Alice permits eBooking services to utilise her email address for sending confirmations, newsletters, and for statistical purposes. Additionally, Alice authorises the booking services to share her email address with trusted partners exclusively to engage with registered services that commit to deleting her email address within a month.

POL

The Privacy Option Language (POL) was formulated by Berthold [2013] to establish privacy contracts between data controllers and data subjects, drawing on the principles of financial option contracts and corresponding data disclosure agreements. Its architecture enforces the 'data minimisation' principle by converting privacy contracts into a standard format. This standardised format ensures that variations in contract compositions are normalised, thus providing a consistent semantic structure across contracts.

Within POL, every privacy contract is dedicated to delineating the responsibilities and entitlements concerning data disclosure. Given its origins in the financial sphere, contract constructions in POL primarily revolve around obligations, except when a straightforward formulation of such rule types is impractical. To specify these formulations, POL relies on various modules that are also open to extension. The language delineates key components, including the **syntax** module, alongside modules addressing **personal data**, **purpose**, **observable** values, and **time**. Additionally, it includes semantics modules focusing on **management** and enhancing **human readability**. The syntax module comprises language primitives essential for defining POL contracts in their standard format. These contracts can then integrate with data modules through various data

support structures, ranging from basic attribute-value pairs, e.g., *(eye colour, brown)*, to intricate tree-like data structures. More specifically, the observable module defines comparison and Boolean operators, which are accessible within the contract execution environment, facilitating evaluations related to e.g. data retention periods. The time module can be used to formalise different time restrictions, e.g. event-driven, discrete, or continuous time. Furthermore, semantic modules are utilised for managing changes in observable variables, e.g., when time elapses, and for translating POL contracts into natural language. Listing 2.14 showcases the semantics of POL through a list of contracts examples: (1) Contract $c_{company}$ delineates the immediate usage of personal data a_1 for purpose p_1; (2) c_{user} represents the negation of $c_{companyA}$ and pertains to the user disclosing the data; (3) c_A signifies a contract wherein a company holds the right to utilise data a_A for purpose p_A at time t_A or can opt not to use it at all (represented by the $zero$ variable in the contract instantiation); and (4) c_B denotes a scenario where a company may or may not utilise data a_B for purpose p_B until time t_B and is obligated to delete it after the deadline t_B.

Listing 2.14 POL contracts extracted from Berthold [2011].

```
(1) c_company = data a_1 p_1
(2) c_user = give c_company
(3) c_A = when (at t_A) (data a_A p_A `or' zero)
(4) c_B = until (at t_B) (data a_B p_B `or' zero)
```

The development of this language took place within the PETWeb II project, primarily aimed at tackling societal inquiries within the electronic identifiers domain. The online documentation offers various application scenarios illustrating POL's utilisation.

PPO

The Privacy Preference Ontology (PPO) [Sacco and Passant, 2011a] proposes a framework for expressing users' privacy preferences regarding the restriction or allowance of access to particular RDF statements within a document. This ontology expands upon WAC to determine users' data access rights, which is limited to specifying who can access an entire RDF document, by allowing finely-grained mechanisms for governing users' access to specific data within RDF resources.

PPO's capabilities for imposing access restrictions extend to individual statements, statement groups, and resources, which can be specific subjects or objects within RDF triples. Additionally, it is essential to specify the type of restriction, as users may be granted either read, write, or both access modes to the data. By utilising the designated **hasCondition** property, specific conditions can be established to delineate privacy preferences concerning particular resources, instances of specific classes or properties, or even specific property values. These conditions can then be checked against a SPARQL ASK query containing all the attributes and properties that users must satisfy to allow or deny access to data.

The authors also created a dedicated privacy preference manager [Sacco and Passant, 2011b] based on PPO. The objective was to empower users to articulate their individual privacy preferences and manage data access based on profile attributes such as relationships, interests, or other shared characteristics.

LegalRuleML

LegalRuleML, a rule language tailored to the legal domain, is developed and maintained by the OASIS[38] LegalRuleML Technical Committee and it attained OASIS Standard status in August 2021 [Palmirani et al., 2021]. This XML-schema specification builds upon and extends RuleML [Boley et al., 2017] and incorporates formal features to represent and facilitate reasoning over legal norms, guidelines, and policies. Key attributes of LegalRuleML include the utilisation of multiple semantic annotations for various legal interpretations, deontic operators, temporal rule management and tracking, and a mapping to RDF.

Hence, the fundamental components of a LegalRuleML document encompass **metadata**, **context**, and **statements**. The metadata segment comprises details concerning the **legal source** of the norms, ensuring their linkage with the corresponding legal text statements. Additionally, it includes information about the **actors** and their **roles** in relation to the established rules, the **jurisdiction**, the **authorities** responsible for rule creation, endorsement, and enforcement, as well as temporal parameters defining rule validity. The context element facilitates the expression of varying interpretations of rule sources, which may evolve over time or differ across jurisdictions. It also facilitates the representation of the **association** element, establishing connections between legal sources and rules. The statements segment involves the formalisation of norms, encompassing constitutive and prescriptive statements, as well as override and violation-reparation statements. Constitutive rules encapsulate definitions outlined in legal documents, whereas prescriptive rules encode deontic specifications. Override statements serve to address conflicting rules, while violation and reparation statements formalise penalties for breaches of norms.

Specifically, Palmirani and Governatori [2018] introduced a framework that leverages LegalRuleML, Akoma Ntoso, and the PrOnto ontology (outlined in Section 2.2.2) to model rules and verify compliance with GDPR's *'Conditions applicable to child's consent in relation to information society services'*, described in Article 8 [2016b].

A-PPL

The Accountable Policy Language (A-PPL), developed by Azraoui et al. [2014], originates from the A4Cloud[39] project, aimed at incorporating accountability requirements into the expression of privacy policies. To achieve this aim, A-PPL extends PPL (PrimeLife Policy Language) by integrating considerations for notification protocols, data storage and retention practices, and auditability guidelines. PPL by Ardagna et al. [2009] is an extensible, XACML-based [2013] privacy policy language established in the context of the PrimeLife[40] project – XACML is an OASIS standard for access control policies that has been previously tested to deal with GDPR requirements related to consent [Fatema et al., 2017] and privacy by design [Piras et al., 2019]. The main concepts within PPL for articulating obligations consist of **triggers** and **actions**. Triggers denote events that can undergo filtering based on specific conditions and are linked to an obligation. These triggers are responsible for initiating actions by the data controller, which are executed in accordance with the data subject's permissions. However, neither PPL nor XACML encompass concepts that

[38] OASIS is a non-profit organisation that focuses on open standards for cloud, security and other domains, https://www.oasis-open.org/ (accessed on 18/July/2023).
[39] http://www.a4cloud.eu/ (accessed on 19/July/2023)
[40] http://primelife.ercim.eu/ (accessed on 19/July/2023)

address needs such as representing information concerning data storage and retention restrictions or incorporating auditability conditions to align with personal data protection regulations.

A-PPL incorporated a role attribute identifier and introduced the role of data protection authority to those already present in PPL, namely the data subject, data controller, and data processor. Additionally, two new triggers for permitting or denying access to personal data were incorporated. Duration and location attributes pertaining to specific processing activities are utilised to enforce data retention and storage rules. Furthermore, A-PPL extends the PPL notification system by specifying the recipient and notification type to be dispatched concerning a particular action. To facilitate auditing, A-PPL introduced a trigger to oversee the data controller's activities and gather evidence of data-related occurrences, which are logged along with parameters such as the activity's purpose, timestamp, or the executed processing operation. Listing 2.15 showcases an instance of an A-PPL obligation to inform a data subject in the event of a personal data breach – the $ActionNotify$ element offers a mechanism for notifying data subjects, triggered in instances of policy violations or data loss.

Listing 2.15 A-PPL example adapted from Azraoui et al. [2014].

```
<Obligation>
    <TriggersSet>
        <TriggerOnPolicyViolation/>
        <TriggerOnDataLost/>
    </TriggersSet>
    <ActionNotify>
        <Media>e-mail</Media>
        <Address>data-subject@example.com</Address>
        <Recipients>Data subject</Recipients>
        <Type>Policy Violation</Type>
    </ActionNotify>
</Obligation>
```

P2U

The work presented by Iyilade and Vassileva [2014] in Purpose-To-Use (P2U) draws inspiration from P3P to construct a policy framework facilitating the sharing of user information across various services and data consumers, grounded in the principle of purpose-driven usage. Its primary objective is to furnish a language tailored for secondary data sharing and usage, with an emphasis on safeguarding user privacy. P2U is structured to encompass details regarding the purpose of data sharing, its duration, and, if desired by the user, potential selling price, while also enabling data consumers to engage in negotiations concerning pricing and retention time.

This policy framework entails the interaction among distinct entities: **users** (who own the data), **data consumers** (services requiring the data), **data providers** (services gathering and sharing the data), and **data brokers** (services overseeing consumer and provider activities, including negotiation tasks). Thus, the principal concepts of P2U are **policies**, **purposes**, **retention** restrictions, **data groups**, and their corresponding **data** elements, and the previously mentioned entities. Policies serve as the foundational component of P2U, with each requiring an associated provider, user, and at least one designated purpose of use. In addition, every policy must be assigned a name

Listing 2.16 P2U example adapted from Iyilade and Vassileva [2014].

```
<POLICY discuri=http://mydatawebsite.com/privacy.html
 ↪  name="ShoppingPolicy">
    <PROVIDER name="FoodIntakeApp" provid="p6528m2" />
    <USER name="Jerry" userid="u1030050503050" />
    <PURPOSE name="Shopping Recommendations" puid="102">
        <CONSUMER name="MyShopApp" consid="c10023" />
        <RETENTION period="180" />
        <DATA-GROUP groupid="g090353" negotiable="TRUE">
            <DATA ref="#dailyfoodintake.food" sell="FALSE" />
            <DATA ref="#dailyfoodintake.quantity" sell="FALSE" />
            <DATA ref="#dailyfoodintake.hungerscale" sell="FALSE" />
        </DATA-GROUP>
    </PURPOSE>
</POLICY>
```

and may optionally include an attribute indicating the path to a human-readable policy, as well as the name and identifier of the corresponding data provider and user. Within a P2U policy, multiple purposes for data sharing can be specified, along with details on retention duration, authorised recipients, and the pertinent data involved. Moreover, the data consumer element includes a **name** property which can be designated as 'public' to allow data sharing with any third-party service. The duration of each purpose's retention period should be defined in days, and an optional **negotiable** property, which defaults to false, can be specified (this term can also be applied to the data group element). The data group component comprises one or more data elements, each capable of being assigned an expiry date, which takes precedence over the retention period of the period, and the option to specify an initial price for the data, should the user opt to sell it. Listing 2.16 illustrates an instance of a secondary data sharing P2U policy – the data provider "FoodIntakeApp" wants to share Jerry's data with the data consumer "MyShopApp" for the purpose of shopping recommendations, allowing the consumer to retain the data for a period of 180 days and to negotiate terms with the provider.

Another publication by the same authors [Iyilade and Vassileva, 2013] outlines a scenario where a user permits data sharing among multiple mobile applications. However, this implementation does not mandate data consumers to adhere to user-defined policies nor does it delineate any particular handling protocols for sensitive data.

SPECIAL

The EU H2020 Scalable Policy-awarE linked data arChitecture For prIvacy, trAnsparency and compLiance (SPECIAL) project endeavoured to create technology that aids in navigating the contemporary tension between privacy and Big Data-based technologies. As such, it aimed to furnish tools for data subjects, controllers, and processors, streamlining the management and transparent utilisation of such data. As a result of this project, two vocabularies were developed: the SPL (SPECIAL Usage Policy Language) and the SPLog (SPECIAL Policy Log) vocabularies [Kirrane et al., 2018a].

A SPL usage policy delineates a collection of permissible actions aligned with the consent of the

Listing 2.17 SPL general usage policy adapted from Bonatti et al. [2019].

```
ObjectIntersectionOf(
    ObjectSomeValueFrom( spl:hasData
        ObjectUnionOf( ex:HeartRate svd:Location ))
    ObjectSomeValueFrom( spl:hasProcessing ex:Profiling )
    ObjectSomeValueFrom( spl:hasPurpose ex:Recommendation )
    ObjectSomeValueFrom( spl:hasStorage
        ObjectIntersectionOf(
            ObjectSomeValuesFrom( spl:hasLocation
                ObjectIntersectionOf( svl:OurServers svl:EU ))
            DataSomeValuesFrom( spl:durationInDays
                DatatypeRestriction( xsd:integer
                    xsd:mininclusive "0"^^xsd:integer ))))
    ObjectSomeValueFrom( spl:hasRecipient spl:AnyRecipient ))
```

data subject. To formalise these actions in accordance with GDPR requirements, SPL outlines five fundamental concepts: the **data** subjected to processing, the intended **purpose** of such processing, detailed information of the **processing** operation, information regarding **storage**, and the designated **recipients** of the processing outcomes. The data storage term encompasses the specification of two attributes: the storage location and duration. Therefore, in mathematical terms, the usage policy is represented as a tuple consisting of five elements, each representing an instantiation of the five core classes, thereby defining a permitted activity. Moreover, a composed usage policy can be formulated by joining a set of authorised processing activities. The vocabularies crafted to delineate each concept within the SPL construct draw upon established privacy-related ontologies. For instance, terms related to processing operations[41] are derived from previous ontologies such as ODRL, while data categories[42], recipients[43], purposes[44], storage duration[45], and location[46] are derived from P3P. These taxonomies have the potential for expansion through the introduction of additional sub-classes [Bonatti et al., 2018a]. An example showcasing this extension possibility is illustrated in Listing 2.17, where the terms `HeartRate`, a sub-class of `svd:Health`, `Profiling` a sub-class of the processing term `svpr:Analyze`, and `Recommendation` as a subclass of the purpose `svpu:Marketing`, are introduced. In this example, data concerning heart rate and location are utilised for user profiling with the aim of generating recommendations, while the data is stored indefinitely within the servers of the data controllers situated in the EU and may be disclosed to any recipients.

SPLog was developed to document the processing events associated with the consent actions granted by data subjects. It leverages PROV-O [Lebo et al., 2013] to incorporate provenance information into the log, aligning with the terminology established for the SPL vocabulary. The key concepts outlined by SPLog encompass the **log** itself and the corresponding **log entries**. Each log is accompanied by metadata, including the software agent to which it pertains, while log entries

[41] https://specialprivacy.ercim.eu/vocabs/processing# (accessed on 20/July/2023)
[42] https://specialprivacy.ercim.eu/vocabs/data# (accessed on 20/July/2023)
[43] https://specialprivacy.ercim.eu/vocabs/recipients# (accessed on 20/July/2023)
[44] https://specialprivacy.ercim.eu/vocabs/purposes# (accessed on 20/July/2023)
[45] https://specialprivacy.ercim.eu/vocabs/duration# (accessed on 20/July/2023)
[46] https://specialprivacy.ercim.eu/vocabs/locations# (accessed on 20/July/2023)

provide details about individual events. These entries can be categorised into two types: policy entries, which are linked to consent forms and associated terms, and data events, e.g., processing or sharing activities. Additionally, these entries should encompass information regarding the involved data subject, event description, content, timestamps, and relevant datasets, facilitating the tracking of event provenance. Hence, SPLog utilises the SPL vocabulary to instantiate the content of a log entry, which allows event grouping to enhance scalability [Kirrane et al., 2018b].

The SPECIAL framework found application across diverse sectors through various use cases: collaborating with *Proximus*[47] to develop personalised tourist recommendations; partnering with *Deutsche Telekom*[48] to deliver traffic alert notifications; and working alongside *Thomson Reuters Limited*[49] to address anti-money laundering requirements.

DPF

The Declarative Policy Framework (DPF), as documented by Martiny et al. [2018] and Martiny and Denker [2020], was developed as part of the DARPA Brandeis program[50]. Its primary objective is to furnish a privacy policy framework grounded in ontology engineering principles and a formal theory of shareability. DPF's policy engine utilises the ontology to delineate policy instantiations, which subsequently inform the generation of user interfaces. These interfaces are designed to empower non-technical users to generate, validate, and manage privacy policies, alleviating them from the intricacies of technical policy language formalisms. Furthermore, DPF's engine is adaptable for integration into systems that support data request management and other Privacy Enhancing Technologies (PETs).

Thus, DPF employs a predefined ontology as a common data model to articulate a specific domain, facilitating the formulation of both permissive and prohibitive privacy policies. Each policy rule encompasses either an allowance or denial statement, necessitating an identifier, a description, an authority, designated data requesters, and the pertinent data affected by the policy, along with the timeframe of its effectiveness. In instances of permissive statements, there is the option to outline a set of constraints that dictate the circumstances under which data may be shared. The policy authority is responsible for assessing whether a given data request aligns with the established policies. Consequently, each data request must include not only the requested data but also the consulted policy authority tasked with granting or denying access, as well as the request timestamp. Subsequently, the request travels through the policy engine pipeline, and upon encountering a matching rule, the engine furnishes the decision along with the identifier and description of the corresponding rule. If the request is authorised, the engine also provides the valid conditions under which it is permissible. Given that a single request may trigger multiple policy rules, the engine must effectively manage conflicting decisions. To address this, DPF incorporates baseline policies, and exceptions are established to delineate policy rules with higher priority concerning the shared data. Through this mechanism, this privacy framework is capable of overriding decisions based on specific constraints.

[47] https://www.proximus.be/ (accessed on 20/July/2023)
[48] https://www.telekom.com/en (accessed on 20/July/2023)
[49] https://www.thomsonreuters.com (accessed on 20/July/2023)
[50] https://www.darpa.mil/program/brandeis (accessed on 20/July/2023)

The ontologies are specified in OWL and can be converted to Flora[51], an object-oriented reasoning system. To demonstrate this framework, the authors offer a pandemic use case wherein national and community policy authorities establish data-sharing policies concerning their residents' health statuses to monitor disease outbreaks. In Listing 2.18, an example DPF policy rule is provided based on this scenario. Any national policy authority, $?pa$, permits nations to share information regarding their residents' disease states, $?reqData$, with response coordinators, $?requester$, at a specified time, $?time$, subject to certain constraints, $?constr$. The $?polData$ query delineates the relationship between nations and the medical statuses of their residents, constrained by $?constr$, which is attached to the $?Resident$ variable. Within this constraint, the birthday of the resident is considered to exclude residents younger than thirteen from the requested data.

Listing 2.18 DPF constrained policy rule adapted from Martiny et al. [2018].

```
@!{NationsAllowConstrainedDiseaseStatesToRCs}
?pa [allow_sa(?requester, ?reqData, ?time, ?constr, ?id, ?descr, 0)] :-
    ?id = "NationsAllowConstrainedDiseaseStatesToRCs"^^\string,
    ?descr = "Nations share disease states w Response
     ↪ Coordinators"^^\string,
    ?pa : NationPolicyAuthority,
    ?requester : ResponseCoordinator,
    ?polData = ${ ?pa [nation -> ?Nation],
        ?Nation : Nation [community -> ?Community, name -> ?NationName],
        ?Community : Community [resident -> ?Resident],
        ?Resident : Person [medicalInformation -> ?MedInfo],
        ?MedInfo : DiseaseStatus [state -> ?MedState],
        ?Resident [constraints -> ?constr] },
    ?thirteenYears is 13*365*24*60*60,
    ?time [subtractTime(?thirteenYears) -> ?latestTime],
    ?constr = [${
        ?Resident : Person [birthDate -> ?Birthdate],
        timeBefore(?Birthdate, ?latestTime) }],
    implies_sharing(?polData, ?reqData, ?constr).
```

LPL

The Layered Privacy Language (LPL), as developed by Gerl et al. [2018], is a privacy language designed to be comprehensible by both humans and machines. Its primary objective is to facilitate the expression and enforcement of GDPR's requirements pertaining to data subject consent, personal data provenance, retention, and the implementation of privacy-preserving processing activities utilising advanced anonymisation techniques. In subsequent research by Gerl and Pohl [2018], efforts were directed towards enhancing LPL to comprehensively address the requirements outlined in Articles 12 to 14 of the GDPR, collectively known as the data subject's 'Right to be informed'.

The policy structure of LPL is organised around purposes. In this structure, a collection of **purposes** forms the core architecture, with each purpose being associated with a set of processed **data** types

[51] http://flora.sourceforge.net/ (accessed on 20/July/2023)

and their corresponding **recipients**. Moreover, the purpose element can be enriched with a human-readable description and also includes properties such as **required**, which specifies whether a particular purpose needs explicit consent from the data subject, and **optOut**, indicating whether the user must actively accept or decline the purpose. Data elements serve to identify the data group to which the processed data belongs, as well as categorise them as sensitive or explicit. Alongside data recipients, other entities like data controllers or the data protection officer can also be designated with LPL. Furthermore, LPL policies can include details on the retention period, data subject's rights, legal basis, and description details pertaining to automated decision-making activities. The example LPL policy provided in Listing 2.19 illustrates how company dr_{C1} operates its personal data handling activities under the LPL privacy policy $lpp_{ds_{U1}-dr_{C1}}$. This policy governs the collection and usage of personal information from a user ds_{U1} for a specific purpose p_{U1}. Additionally, it allows for the optional sharing of collected data with a third-party recipient. Should such sharing occur, a new contract must be established, with company $C1$ (ds_{C1}) acting as the data source and the third party $C2$ (dr_{C2}) as the data recipient. The purpose of processing, denoted as p_{U1}, is 'Marketing', involving personal data $\hat{D}_1$ such as postal code (anonymised via the 'Suppression' method) and salary information, which must be deleted within 180 days following the fulfilment of the purpose.

Listing 2.19 LPL policy extracted from Gerl et al. [2018].

1 ds_{U1}=('U1','Person',$publicKey_{U1}$,'DataSource')
2 dr_{C1}=('C1','Legal Entity',$publicKey_{C1}$,'DataRecipient')
3 ds_{C1}=('C1','Legal Entity',$publicKey_{C1}$,'DataSource')
4 dr_{C2}=('C2','Legal Entity',$publicKey_{C2}$,'DataRecipient')
5 $lpp_{ds_{U1}-dr_{C1}}$=('1','LPP1','en','https://company.com/privacy.html',$\varnothing$,ds_{U1},{p_{U1}})
6 p_{U1}=('Marketing','false','true','Marketing purposes, including newsletters.',{dr_{C1},dr_{C2}},r_1,pm,$\hat{D}_1$)
7 r_1=('AfterPurpose','180 days')
8 $\hat{D}_1$={d_{postal},d_{salary}}
9 d_{postal}=('postal-code',dGroup,'Number','true','Postal code of the user','QID',am1)
10 am_1=('Suppression',{ama_1,ama_2,ama_3,ama_4},$\varnothing$)
11 ama_1=('Suppression Replacement','*')
12 ama_2=('Suppression Direction','backward')
13 ama_3=('Minimum Level','2')
14 ama_4=('Maximum Level','4')
15 d_{salary}=('salary',dGroup,'Number','true','Monthly salary amount received by the user','Sensitive',$\varnothing$)

Gerl and Meier [2019] conducted validation of this language through a real-world use-case scenario within the healthcare domain, showcasing its effectiveness and constraints concerning GDPR compliance. Subsequent research expanded upon LPL by integrating machine-readable privacy icons [Gerl, 2018], aiming to evaluate their impact on the comprehension of privacy policies. Additionally, an LPL Personal Privacy Policy User Interface was introduced [Gerl and Prey, 2018]. This interface primarily aims to present information pertinent to privacy policies, aiding data subjects in providing informed consent. It includes a policy header containing a link to the human-readable policy and an overview of processing purposes using the previously mentioned privacy icons. Furthermore, a purpose section outlines all purposes outlined in the privacy policy, along

with details regarding data controllers, recipients, retention periods, and anonymisation methods.

2.3.3 Comparative analysis

Using Table 2.7, it is possible to assess and compare the policy languages outlined in this Section, regarding their effectiveness in aiding the representation of GDPR rights and obligations. The languages in the Table are arranged firstly by the number of supported criteria in descending order, followed by alphabetical sorting if needed, to enhance clarity and readability.

Table 2.7: Comparison of the analysed privacy policy languages, according to the defined criteria described in Section 2.3.1.

	C1	C2	C3	C4	C5	C6
LegalRuleML	Yes	Partially	No	Yes	Yes	Yes
ODRL	Yes	Partially	Yes	No	Yes	Yes
SPL	No	Partially	Yes	Yes	No	Yes
A-PPL	Yes	Partially	No	Yes	No	No
DPF	Yes	Partially	No	Yes	Unknown	No
P3P	No	Partially	Yes	No	No	Yes
AIR	No	No	No	Yes	No	Yes
LPL	No	Partially	No	Yes	Unknown	No
S4P	No	Partially	No	Yes	No	No
P2U	No	Partially	No	No	No	No
POL	No	Partially	No	No	No	No
PPO	No	No	No	No	No	No
XPref	No	No	No	No	No	No

While these languages may not explicitly address the rights and obligations outlined in Section 1.3.4, they can still encapsulate some of the information items discussed therein. Thus, they are categorised as partially capable of representing GDPR concepts and principles (C2 criterion in Table 2.7). Most of the examined languages can partially fulfil the representational requirements of GDPR as identified in Table 2.2, with the exceptions being AIR, PPO, and XPref. Examples illustrating how to encode specific aspects of privacy policies for each language partially capable of representing GDPR concepts are provided in Listings 2.11 to 2.19.

Among these languages, only ODRL, SPL, and P3P offer taxonomies for populating policies. Moreover, only LegalRuleML, ODRL, A-PPL, and DPF incorporate deontic concepts like permissions or obligations into their models. In their documentation, LegalRuleML, SPL, A-PPL, DPF, AIR, LPL, and S4P also acknowledge the presence of reasoning mechanisms or other supportive tools, which leverage the implemented languages to aid in compliance efforts. Some of these languages also provide access to such tools. Nevertheless, LegalRuleML and ODRL are the only languages that are actively maintained and developed, while, among the languages examined, only LegalRuleML, ODRL, SPL, P3P, and AIR offer resources that can be readily reused on the Web.

Notably, LegalRuleML and ODRL distinguish themselves from other languages by possessing the capabilities to positively address a larger proportion of the established comparison criteria,

including the ability to model deontic concepts and GDPR terms, e.g., purposes or data recipients. Moreover, their development and extension can be supported by the community groups in charge of their maintenance, and their resources are open and accessible. Beyond this, ODRL also supports the modelling of other constraints with particular importance in the definition of access control policies, e.g., spatial and temporal constraints, the representation of distinct types of policies, e.g., offers, requests and agreements, and has a profile mechanism to develop extensions to its core vocabulary. As such, ODRL will be the basis upon which access policies are expressed in this Thesis.

2.4 Gaps and challenges

This Section discusses the challenges of having a transparent, legally-aligned Solid ecosystem based on the literature review described in this Chapter. Firstly, in Section 2.1, the need to have proper identification and separation of roles in decentralised data environments, in particular for accountability purposes, as well as of information regarding the infrastructure used for data storage. Moreover, the issue related to the availability and discovery of particular types of data still has challenges to be solved if interoperability is to be achieved. Additionally, the specification of machine-readable and machine-actionable data subjects' privacy preferences and of data controllers' data handling practices, as well as the expression of provenance metadata, e.g., related to data subjects' rights exercising, is not up-to-date with current EU data protection law requirements, as concluded in Sections 2.2 and 2.3. As such, said gaps must be closed in order to allow legal compliance checks, e.g., by supervisory authorities. Beyond the technical and legal needs aforementioned, societal and business demands should also be considered as the described decentralised environments are not used in isolation by data subjects and will be an important component for an open and diverse data economy. Based on these performed analysis, the identified gaps can be translated into the following issues:

Ch1. **Identity of Solid actors and their roles is unknown** – Solid users lack awareness of the entities responsible for providing and/or developing their Pods, the applications they utilise, WebIDs, or other server infrastructure. Additionally, the majority of applications or services fail to give contact details or information about their data protection officer.

Ch2. **No metadata about Solid infrastructure** – Solid users lack information regarding the Solid specification their Pod is operating on, the services installed within it, or the location of the servers where Pods are hosted. Additionally, there is no record of this information kept in the Pod for convenient reference by the user.[52]

Ch3. **Availability/Discovery of categories of data** – In order for Solid applications to access data within Pods at a granular level, i.e., by data type, they require knowledge of its existence and storage location within the Pod. Additionally, for smooth interoperability, it is essential to document the schemas, formats, or shapes for data recognised or supported by applications, services, or Pods.

[52] An incomplete catalogue of Pod providers is published at https://solidproject.org/users/get-a-pod, detailing the hosting service used for the Pods (although no specifics on the entities behind them are provided) and, occasionally, the country of hosting (though lacking a privacy policy for the storage service).

Ch4. **Pod and applications providers do not provide information on their data processing practices** – The majority of providers and developers offering Pod-related services fail to give human and/or machine-readable privacy notices or specify the data they require for operation. It is imperative to document this information within the Pod itself to enable users to retain a record of data requests. This ensures that users have a reference in case data is utilised in a manner not authorised by them.

Ch5. **Users cannot express their privacy policies** – Solid users lack the means to articulate their privacy preferences and requirements, as well as to oversee incoming data requests or manage existing agreements regarding data usage.

Ch6. **No logging or record-keeping** – There is no recorded provenance metadata in the user Pod for accountability purposes. For instance, users do not maintain consent records or information about who has accessed their data, how it is being utilised, or any alterations to data policies.

Ch7. **No legal compliance checks** – Currently there are no Solid-based tools for Solid users to address legal obligations, like granting/revoking consent or exercising rights under the GDPR. Additionally, no tools are available for the authorities conducting investigations to access required auditing information.

Ch8. **Societal and business needs** – Beyond technical and legal requirements, user studies still need to be performed to understand what type of policies fulfil the needs and expectations of users. Furthermore, a similar exercise needs to be performed for companies in order to understand how they can function and adapt their business for such decentralised data-sharing environments.

The following Chapters of this Thesis will tackle these challenges in distinct manners, with a particular focus on the representation of legally-aligned information for the access to subject's data stored in decentralised data environments.

Chapter 3

Objectives and Contributions

3.1 Objectives

To address the challenges and limitations identified in Section 2.4, the main objective of this Thesis is the following:

> Research methodologies and design vocabularies and services to aid EU data subjects in taking control of the movement of their personal data.

This main objective is divided into the following three sub-objectives:

O1. Design methods and systems to assist data subjects (in representing their privacy preferences and consent) and data controllers (with GDPR requirements) in order to support automated data transactions, accountability and transparency.

O2. Design a policy matching algorithm that utilises the developed vocabularies to express data-sharing preferences, requests and agreements in decentralised personal datastores.

O3. Design a service, using state of the art vocabularies, to assist with representing information connected with GDPR's data subject rights.

3.2 Hypotheses

H1. The use and extension of data protection vocabularies and machine-readable policy languages is suitable for the representation of consent terms and fine-grained policies for the processing of personal data.

H2a. Data protection vocabularies and policy languages can be used to establish fine-grained access control conditions to personal data.

H2b. Semantic Web vocabularies can be used to describe metadata related to decentralised personal datastores, including entities, infrastructure and roles.

H2c. Data protection vocabularies can be used to represent machine-readable information related

Figure 3.1: Overview of the objectives and contributions of this Thesis.

to data subject's rights.

H3. Semantic Web vocabularies and decentralised technologies can be used as a basis for establishing a policy matching process for the achievement of data-sharing agreements which fulfil the data subjects privacy preferences.

3.3 Assumptions

The work presented in this Thesis is done under the following set of assumptions:

A1. Decentralised personal datastores function as a personal information management system and as such should be regulated according to personal data processing law, whereas broader datastores can be used for storing all types of data and therefore imply a myriad of additional legislation, which is not in the scope of this research.

A2. It is viable and advantageous for users to manage their data and privacy preferences through decentralised services and applications.

3.4 Restrictions

The work presented in this Thesis is subject to the following restrictions:

R1. The scope of the research will be restricted to decentralised datastores based on the technology stack of the Semantic Web.

R2. The research in this Thesis is focused on the personal data protection domain in the European Union as it presents a fully-fledged legal regime, the GDPR, which puts the data subjects at the centre of the flow of their personal data.

R3. The legal analysis does not contemplate the Data Protection Law Enforcement Directive 2016/680, which was launched together with the GDPR.

R4. The research in this Thesis is restricted to personal data that is digitally available in an open format, e.g. RDF, CSV, or PNG files.

R5. The research in this Thesis focused on determining access control to decentralised data – usage control, i.e., what happens to the data once it has been accessed, is out of the scope.

3.5 Research questions

The main research question that supports this Thesis is:

> To what extent are Semantic Web vocabularies and decentralised technologies able to support the exercising of data subject rights and determine the access conditions to personal data?

We divide the main research question into the following sub-questions:

RQ1. Are Semantic Web standards and specifications able to represent information related to privacy preferences, data access policies, and other metadata, aligned with personal data protection requirements?

RQ2. Can Semantic Web vocabularies be used to determine access control conditions to personal data stored in decentralised data systems?

RQ3. Is it possible, using decentralised Web technologies, to facilitate the exercising of data subject rights in light of the GDPR?

3.6 Contributions

The contributions of this Thesis are outlined below. In order to centralise their access, a Web page with links to all contributions is available at https://w3id.org/people/besteves/phd/contributions, including links to open-access versions of the publications mentioned in Section 1.4.

3.6.1 Main contributions

C1. Development of Vocabularies

C1.1. OAC: Development of an ODRL Profile for Access Control (OAC), to define access control policies that express permissions and/or prohibitions associated with data stored in a decentralised storage environment, such as Solid Pods.

C1.2. PLASMA: Development of a Policy LAnguage for Solid's Metadata-based Access control (PLASMA), to provide consistent taxonomies to describe the entities, infrastructure, legal roles, policies, notices, registries, and logs necessary to understand and establish responsibilities and accountability within the Solid ecosystem.

C1.3. Rights Exercising: Development of vocabulary-based patterns to describe rights exercising metadata using DPV, to provide uniform recording of data subject rights exercising activities.

C1.4. DUODRL: Development of ODRL rules for the Data Use Ontology (DUO), to create policies for the sharing of health data.

C1.5. DGAterms: Development of a Data Governance Act (DGA) vocabulary, to create OAC-based policies for the sharing of data for altruistic purposes and keep registries of available datasets.

C2. Policy matching algorithm: Design and implementation of a policy matching algorithm and data-sharing agreement generator prototype for access to data stored in Solid Pods.

3.6.2 Secondary contributions

C3. Analysis of data protection-related challenges for decentralised datastores: A complete literature review was performed for existing work on Solid, machine-readable policy

languages and data protection vocabularies in Chapter 2. From this review, a series of gaps and challenges in the literature were identified and described in Section 2.4.

C4. Proof of concept prototypes

C4.1. SOPE: Development of the Solid ODRL access control Policies Editor (SOPE), to generate and store OAC policies in Solid Pods.

C4.2. SoDA: Development of a Solid Data Altruism application (SoDA), to implement data altruism as a service using the Solid protocol and ODRL policies to grant access to personal data for altruistic purposes in a privacy-friendly manner.

C4.3. Service for exercising data subject rights: Design and implementation of a service to generate rights exercising metadata.

C4.4. Service to search for data protection-related concepts: A REST API service to find references to specific concepts in the collection of identified ontologies and languages[1].

3.6.3 Contributions to W3C Community Groups

C5. Contributions to W3C DPVCG: When aligned with the groups' purpose of having metadata to describe personal data handling activities, the concepts present in the developed vocabularies were submitted for integration in the DPV's specifications. Contributions to the DPV primer were also submitted.

C6. Contributions to W3C ODRL CG: OAC was submitted to be considered as an official ODRL profile for Access Control. The ODRL-related work developed in this Thesis was also considered for the under-development specification of a formal semantics document for ODRL.

3.7 Research Methodology

The research work for this Thesis involves two distinct knowledge domains: law and ontology engineering. As such, distinct research methods were followed for the distinct stages of this research, in particular:

- To analyse knowledge sources (Section 3.7.1)
- To review state of the art solutions to represent privacy terms in decentralised settings (Section 3.7.1)
- To create and validate vocabularies (Section 3.7.2)
- To publish and archive research software (Section 3.7.3)

[1]Available at https://w3id.org/people/besteves/phd/sota/searcher.

3.7.1 Literature review

In this Section, the methodology to analyse legal knowledge sources and review state of the art solutions to represent privacy terms in decentralised settings is introduced.

Review of legal knowledge sources

The legal knowledge, upon which this research is based, is derived from the General Data Protection Regulation. In particular, an analysis was made of Chapters III and IV ('Rights of the data subject' and 'Controller and processor', respectively), where each article in both chapters was manually studied to search for interactions between the entities and the information that needs to be exchanged between them.

In addition to the text of the GDPR, the following sources were utilised or mentioned:

- Guidelines and opinions published by EDPB[2].
- Joint opinions and technical reports published by EDPS[3].
- Guidelines and opinions published by the Article 29 Data Protection Working Party (WP 29)[4].
- Other data-related legislation of the European Parliament and Council, in particular, the DGA, the eIDAS and its proposed amendment, and the EHDS proposal.
- Research publications in the personal data protection domain, in particular, related to the GDPR.

Literature Review of Solid, Policy Languages and Data Protection Vocabularies

Throughout the years different methodologies have been published for conducting a literature review [Webster and Watson, 2002, Kitchenham and Brereton, 2013] and, in particular, in 2019, a survey of distinct categories of methodologies, including guidance on how to execute and evaluate them, was published by Snyder [2019]. Three types of review methodologies are described, namely, systematic, semi-systematic, and integrative approaches. Also according to Snyder, the choice of the correct approach is related to the research questions, purpose, or style of the document being reviewed.

As such, an integrative approach [Whittemore and Knafl, 2005] was used as it is the most appropriate for the objective of synthesising academic publications, in particular regarding the description and development of different policy languages and data protection vocabularies in a qualitative and quantitative manner, i.e., coverage of distinct personal data-related concepts and count of modelled concepts, respectively. The same approach was taken to evaluate research on semantic-based personal datastores. Moreover, the snowballing procedure [Wohlin, 2014] was followed to search for relevant academic publications to be included in this Thesis, as well as other public documentation as it is advised by the integrative literature review approach. Additional citation

[2] https://edpb.europa.eu/ (accessed on 22 October 2023)
[3] https://edps.europa.eu/ (accessed on 22 October 2023)
[4] https://ec.europa.eu/newsroom/article29/items/itemType/1360 (accessed on 22 October 2023)

analysis research was performed using Webster and Watson's backward and forward snowballing methodologies – examine the reference list of the already identified publications to determine new documents that should be considered and select new research publications that cite the ones already being considered, respectively.

The collected publications were reviewed, evaluated, and, if relevant, included in the state of the art, according to the following criteria:

- Availability of a publication to review.
- Only publications in English were considered.
- Existence of online resources, e.g., ontology documentation or RDF/OWL specifications, was considered beneficial, as it allows for a better understanding and a quantitative evaluation of the reviewed solution.
- Pre- and Post-GDPR works were considered.

3.7.2 Ontology engineering

The Linked Open Terms (LOT)[5] methodology for the specification, implementation, publication, and maintenance of ontologies [Poveda-Villalón et al., 2022] was used for the development of the ontologies in this Thesis, as it is based on existing methodologies for the development of Semantic Web technologies and it is aligned with software development and research projects lifecycles. Moreover, the Suárez-Figueroa et al. [2012] methodology was used to define formal competency questions (CQ) and to collect the ontologies requirements, which are then consolidated in an Ontology Requirement Specification Document (ORSD). Figure 3.2 presents a diagram of the main steps of the LOT methodology workflow, with a specific focus on the ontologies implementation phase.

Taking into account the requirements specified in the ontologies' ORSD, their first conceptualisations were then generated through a visualisation tool, the Chowlk Visual Notation tool[6] [Chávez-Feria et al., 2022], followed by feedback from experts. The generated conceptualisation diagrams were then used to generate the first version of the ontologies encoding as a Turtle file. Furthermore, after the generation of the first version of the ontologies, the created terms were evaluated against a set of use case scenarios and using SPARQL queries. From this evaluation, if necessary, new concepts were added to the ontologies, and the ORSDs were also updated accordingly. The ontologies were also evaluated by using the OntOlogy Pitfall Scanner (OOPS!)[7] [Poveda-Villalón et al., 2014] to detect common errors in ontology development, such as missing domain or range properties or missing annotations, and using FOOPS![8], the Ontology Pitfall Scanner to ensure ontology alignment with the Findable, Accessible, Interoperable and Reusable (FAIR) principles [Garijo et al., 2021]. As a final evaluation, ontologies classes and properties and, in particular, their definitions

[5] More details regarding the methodology and the tools promoted by LOT are available at https://lot.linkeddata.es/ (accessed on 14 June 2023).

[6] The Chowlk Converter tool and respective usage instructions are available at https://chowlk.linkeddata.es/ (accessed on 14 June 2023).

[7] The OOPS! tool is available at https://oops.linkeddata.es/ (accessed on 14 June 2023).

[8] The FOOPS! tool is available at https://w3id.org/foops (accessed on 30 November 2023).

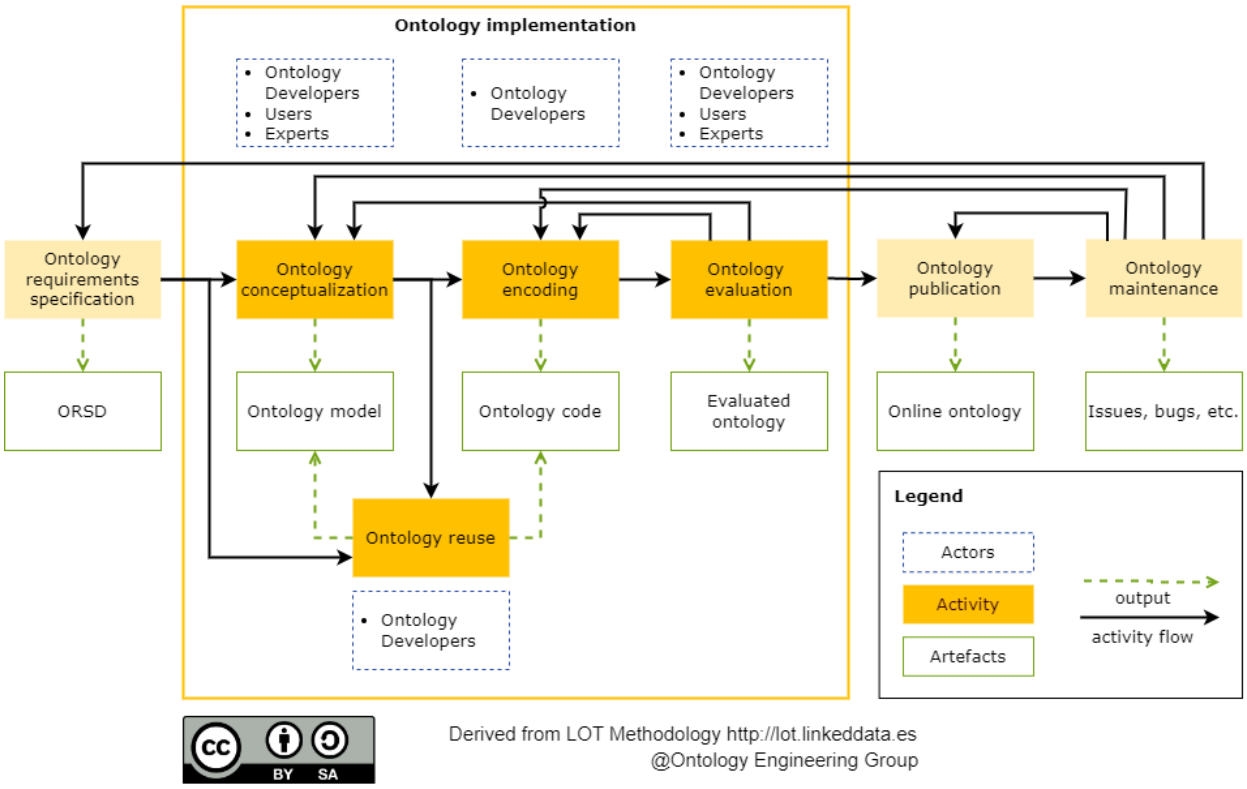


Figure 3.2: Ontology development workflow based on the LOT methodology.

were reviewed by legal and technical experts and, when applicable, connected with the relevant legal rules, guidelines, and other literature.

The ontologies are published using the `w3id.org`, "Permanent Identifiers for the Web", service[9]. This service provides a secure and permanent re-direction and content negotiation service that serves both human-readable documentation and a machine-readable file from the same URI. The source code is hosted on GitHub[10] and version control is done using Git[11].

3.7.3 Publication and archival of research software

Recently, the scientific community has also been discussing extending FAIR data practices to research software [Martinez et al., 2019, Gruenpeter et al., 2024]. Extending such practices to software implies the inclusion of rich metadata, in machine and human-readable format, and unique persistent identifiers for software to be findable and accessible. In terms of interoperability and reusability, software must include clear documentation and reuse instructions, as well as use common standards and platforms. As such, the following best practices, adopted from the aforementioned guidelines, are followed to ensure the FAIR publication and preservation of research software in this Thesis:

[9]This service is run by the W3C Permanent Identifier Community Group (`https://www.w3.org/community/perma-id/`, accessed on 22 October 2023).
[10]`https://github.com/` (accessed on 22 October 2023)
[11]`https://git-scm.com/` (accessed on 22 October 2023)

1. Description of the software is included in the README.md file of the software repository.
2. Software is archived in a software registry, i.e., Zenodo.
3. Software has a persistent identifier such as Digital Object Identifier (DOI) and/or a w3id.org.
4. Software is downloadable.
5. Software follows a semantic versioning scheme.
6. Software requirements are listed in the software repository.
7. Software installation instructions are included in the README.md file of the software repository.
8. Software usage instructions are included in the README.md file of the software repository.
9. The software repository has a license.
10. The software repository provides instructions on how to cite it.
11. The software repository includes metadata including programming language, creation date, keywords, and releases.

Moreover, the source code of the developed research software is hosted on GitHub and version control is done using Git.

3.8 Evaluation Methodology

As previously mentioned, the work for this Thesis involves two distinct research fields, law and ontology engineering. As such, distinct evaluation methods were followed to assess the hypotheses identified in Section 3.2, in particular:

E1. (for H1.) The goal of this evaluation is to determine the alignment of the developed models to represent consent terms and fine-grained policies for the processing of personal data with the EU's General Data Protection Regulation. To this end, the proposed vocabularies were validated by legal experts through collaboration with members of W3C DPVCG and legal scholars from the PROTECT ITN. Moreover, alignment with the ISO/IEC 27560 standard on consent records and receipts was also verified.

E2. (for H2a., H2b., and H2c.) The goal of this evaluation is to assess whether the developed methods can be used to establish access control conditions, describe metadata related to decentralised personal datastores, and represent information related to data subject's rights. To this end, the quality of the proposed ontologies was evaluated by detecting common pitfalls and alignment with FAIR principles, and their ability to answer the competency questions through SPARQL queries was also verified.

E3. (for H3.) The goal of this evaluation is to test whether the developed ontologies and policy-based algorithms can be used to define data access agreements that fulfil the data subjects' privacy preferences. To this end, a proof of concept implementation for policy matching

towards the generation of data access agreements was built to evaluate the proposed algorithms against a specific real-world use case involving health data sharing. Furthermore, proof of concept prototypes to generate policies and exercise the GDPR's right of access were also built, to assess the applicability of the developed vocabularies in the development of decentralised applications, in addition to a data altruism protocol which verifies the extensibility of the proposed vocabularies to cover other data protection laws, e.g., the EU's Data Governance Act.

Part II

GDPR-ALIGNED VOCABULARIES FOR PERSONAL DATASTORES

Chapter 4

Vocabularies for Personal Datastores

The content of this Chapter has already been partially included in the articles published during this Thesis [Esteves et al., 2021, Esteves and Pandit, 2023, Esteves et al., 2022a].

The source code produced during the development of this chapter is stored at:

- https://w3id.org/oac/repo
- https://w3id.org/oac/policies
- https://w3id.org/plasma/repo
- https://w3id.org/people/besteves/justifications/repo
- https://w3id.org/people/besteves/rights/repo

This Chapter builds upon existing Semantic Web standards and specifications to develop a set of vocabularies that can support data subjects in the expression of their privacy preferences when it comes to accessing their data and exercising their rights, as well as data controllers to deal with their *"[t]ransparent information"* requirements, explicitly set in GDPR's Articles 12–14. As previously established in Section 2.1, Solid's access control and interoperability specifications do not contain the terms to satisfy said requirements, and as such, the incorporation of these vocabularies will lead to a GDPR-aligned personal datastore.

Thus, Section 4.2 describes the development of an ODRL profile (OAC) with the main goal of defining legally aligned policies that express permissions and/or prohibitions associated with purpose-based access to data stored in decentralised storage environments, such as Solid Pods. Such policies will be used to express the data subjects' preferences concerning the access to their *personal* data, to represent requests to access data, and to record the agreed access conditions for future inspection.

Section 4.3 describes the development of a metadata language for Solid (PLASMA) to provide consistent taxonomies to describe the entities, infrastructure, policies, notices, registries, and logs necessary to understand and establish responsibilities and accountability within the Solid ecosystem. PLASMA utilises OAC to express data policies, provides a set of conformance conditions that should be met by Pod, app, and service providers, as well as users and agents, to comply with

the established specification and a description of workflows where PLASMA terms should be used to satisfy such conformance conditions.

Section 4.4 showcases the usage of vocabulary-based, e.g. DPV, DCMI, PROV-O, and DCAT, patterns to describe rights exercising metadata with the goal of providing uniform records of data subject rights exercising activities.

Section 4.5 presents the results of the ontologies evaluation, including the detection of common pitfalls with OOPS!, alignment with FAIR principles with FOOPS! and validation of competency questions with SPARQL queries, and Section 4.6 discusses the alignment with the ISO/IEC 27560 standard on 'Consent record information structure'.

The methodology followed to develop and evaluate the vocabularies described in this Chapter is described in Section 3.7.2. The prefixes and namespaces used in the Listings in this Chapter are explicitly defined in the Namespaces list.

4.1 Background

As established through the state of the art and in the comparative analysis performed in Sections 2.2.3 and 2.3.3, DPV contains the highest number of concepts to model GDPR's rights and obligations and their privacy terms, is being actively developed and maintained, is open and accessible, and ODRL supports the modelling of deontic concepts, e.g., permissions or obligations, constraints, e.g., spatial and temporal, and types of policies, e.g., offers, requests and agreements, and has a mechanism to develop extensions to its vocabulary through profiles. As such, they can be used as a starting point to express policies for access to personal data, while invoking privacy and data protection-specific terms.

Figure 4.1 presents a diagram of the ODRL Information Model. Its main goal is to *"enable flexible Policy expressions by allowing the policy author to include as much, or as little, detail in the Policies"* [Iannella and Villata, 2018], using the terms defined in the ODRL Vocabulary & Expression specification [Iannella et al., 2018]. Table 4.1 provides an overview of the concepts modelled in the ODRL vocabulary.

The model also expresses which properties are mandatory and optional to define policies and their respective entities, assets, actions and constraints. Both recommendations are being promoted and maintained by the W3C ODRL CG, which also aims to support the development of ODRL profiles and publish reports related to ODRL usage, such as:

- the ODRL Implementation Best Practices [Smith et al., 2023], which presents examples of ODRL usage and describes good implementation practices;
- the ODRL Profile Best Practices [Steidl, 2023], which presents guidelines for the development, definition and publication of ODRL Profiles;
- the ODRL Formal Semantics [Fornara et al., 2023], which discusses and provides a formal semantics specification to ensure the correctness and consistency of services that use ODRL.

While ODRL presents itself as a well-tested resource for the expression of policies, it does contain the concepts to model personal data-related access policies or to invoke data protection-related

Table 4.1: Overview of the concepts modelled in the ODRL vocabulary.

Concept	Subclasses
Policy	Agreement, Assertion, Offer, Privacy, Request, Set, Ticket
Rule	Duty, Permission, Prohibition
Party functions	assignee, assigner, attributedParty, attributingParty, compensatedParty, compensatingParty, consentedParty, consentingParty, contractedParty, contractingParty, informedParty, informingParty, trackedParty, trackingParty
Action	Attribution, CommericalUse, DerivativeWorks, Distribution, Notice, Reproduction, ShareAlike, Sharing, SourceCode, acceptTracking, aggregate, annotate, anonymize, archive, attribute, compensate, concurrentUse, delete, derive, digitize, display, distribute, ensureExclusivity, execute, extract, give, grantUse, include, index, inform, install, modify, move, nextPolicy, obtainConsent, play, present, print, read, reproduce, reviewPolicy, sell, shareAlike, stream, synchronize, textToSpeech, transfer, transform, translate, uninstall, use, watermark
Operand	and, andSequence, or, xone
Left Operand	absolutePosition, absoluteSize, absoluteSpatialPosition, absoluteTemporalPosition, count, dateTime, delayPeriod, deliveryChannel, elapsedTime, event, fileFormat, industry, language, media, meteredTime, payAmount, percentage, product, purpose, recipient, relativePosition, relativeSize, relativeSpatialPosition, relativeTemporalPosition, resolution, spatial, spatialCoordinates, systemDevice, timeInterval, unitOfCount, version, virtualLocation
Operator	eq, gt, gteq, hasPart, isA, isAllOf, isAnyOf, isNoneOf, isPartOf, lt, lteq, neq
Right Operand	policyUsage

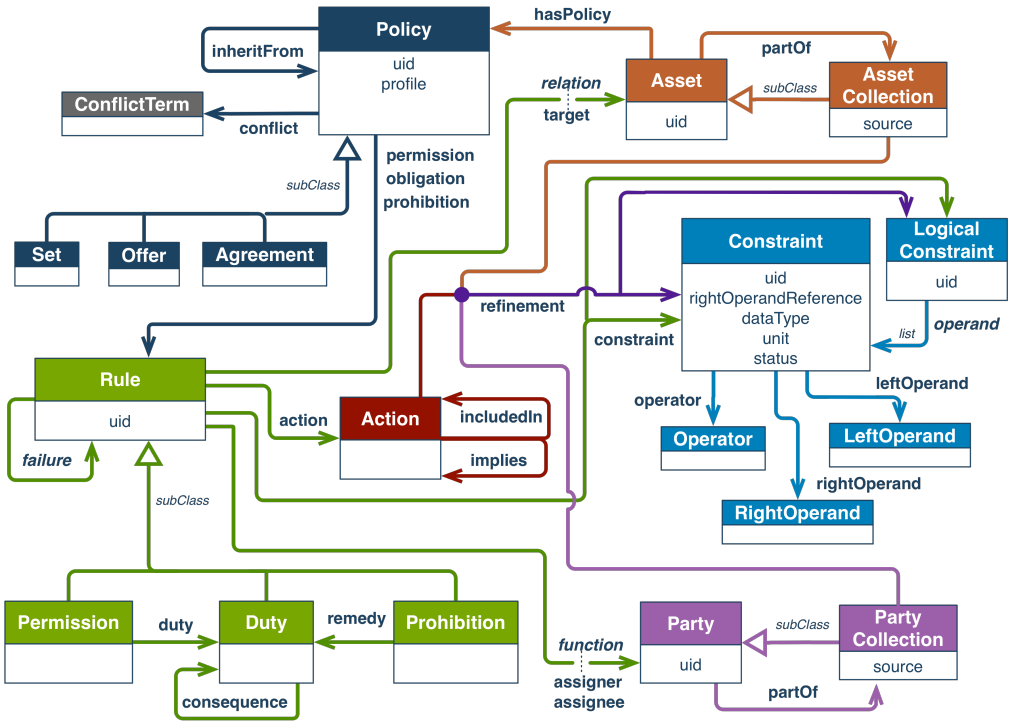


Figure 4.1: ODRL Information Model, adapted from Iannella and Villata [2018].

terms. As such, its profile mechanism provides an opportunity to extend the ODRL vocabulary with these missing terms, e.g., by associating it with personal data-focused vocabularies such as DPV. Figure 4.2 provides an overview of DPV's core concepts.

As indicated in the state of the art Chapter of this Thesis, DPV provides the most extensive list of data protection-related terms among the evaluated solutions. Table 4.2 includes a list of the taxonomies defined in DPV's main specification, as well as the number of classes and properties defined in each taxonomy and, in the third column, the number of classes and properties that were contributed to the vocabulary in the course of the development of this Thesis. These contributions were derived from the work developed and described throughout this Chapter, as well as Chapters 6 and 7.

Moreover, the DPVCG also published a primer document [Pandit et al., 2022], which provides a description of DPV and its concept modelling, examples that illustrate how the provided concepts should be used to represent metadata regarding personal data handling activities and guidelines towards the application of DPV in particular use cases, e.g., consent record keeping or rights exercising. Additionally, as previously described in Section 2.2.2, the DPVCG developed six extensions to the main specification, to model personal data categories, GDPR-specific concepts, technology and jurisdiction-relevant concepts, risk, and EU rights concepts. Table 4.3 includes a list of the DPV's extensions, as well as the number of classes and properties defined in each extension and, in the third column, the number of classes and properties that were contributed to the extensions in the course of the development of this Thesis.

Additionally, existing work, using ODRL's profile mechanism, has been published to instantiate

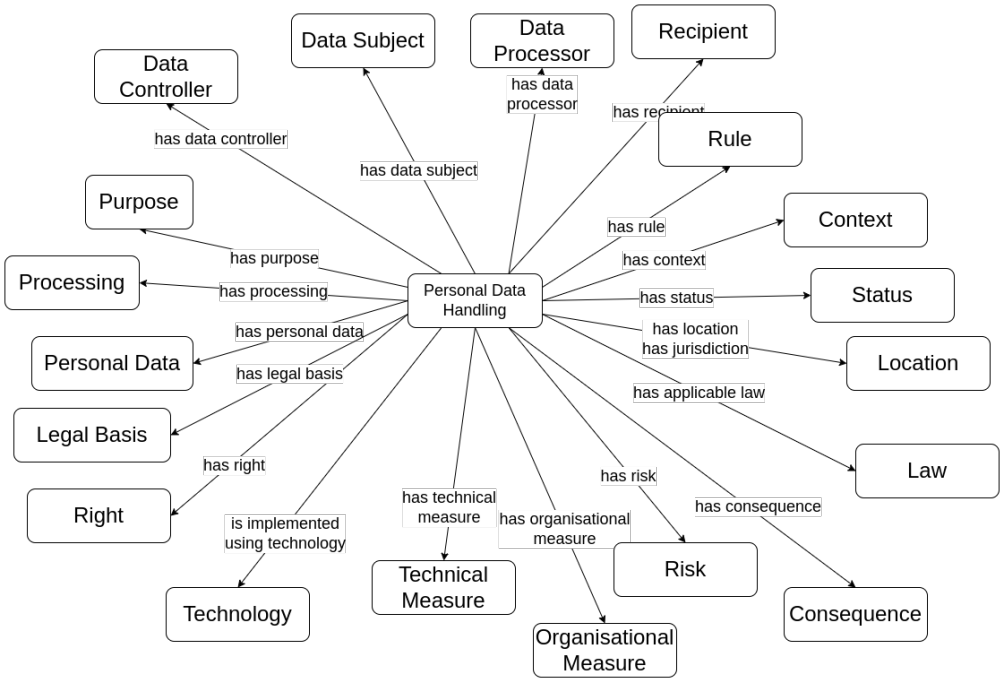


Figure 4.2: Overview of DPV's core concepts, adapted from Pandit et al. [2022].

Table 4.2: Taxonomies defined in DPV's main specification, with the respective number of defined classes and properties, as well as the number of contributions of this Thesis to the vocabulary.

Taxonomies	#Classes (#Properties)	Contributions
Entities	4 (7)	1 (4)
Legal Roles	9 (9)	0 (0)
Authorities	5 (2)	0 (0)
Organisations	9 (0)	0 (0)
Data Subjects	26 (2)	17 (0)
Purposes	78 (2)	20 (0)
Processing	45 (1)	0 (0)
Storage Conditions & Automation	29 (5)	3 (0)
Scale of Processing	27 (4)	0 (0)
Data	16 (2)	0 (0)
TOMs	139 (6)	4 (0)
Legal Bases	34 (5)	0 (0)
Duration & Frequency	23 (11)	7 (3)
Status	39 (5)	0 (0)
Location & Jurisdiction	25 (5)	0 (0)
Risk & Impacts	16 (12)	4 (3)
Rights	9 (2)	9 (0)
Rules	4 (4)	4 (4)

Table 4.3: DPV's extensions with the respective number of defined classes and properties, as well as the number of contributions of this Thesis to the extensions.

Extensions	#Classes (#Properties)	Contributions
Personal data	206 (0)	3 (0)
GDPR	92 (0)	16 (0)
Technology	60 (8)	0 (0)
Jurisdiction	452 (0)	0 (0)
Risk	376 (0)	0 (0)
EU Rights	62 (0)	0 (0)

GDPR Articles as ODRL obligations [Agarwal et al., 2018] and as permissive, prohibitive or obligated policies with dispensations, which are translated into Answer Set Programming (ASP) rules for compliance checking [De Vos et al., 2019]. Other ODRL-based works have been published, related to (i) the representation of agreements to access data and execute algorithms in digital marketplaces [Shakeri et al., 2019], (ii) the dynamic generation of privacy policies for IoT-generated data [Cano-Benito et al., 2023], and (iii) the representation of privacy policies as ODRL requests, which use a small subset of DPV's taxonomies and do not follow the ODRL Information Model [Krasnashchok et al., 2020].

4.2 ODRL profile for Access Control

This Section describes the development of OAC, an ODRL profile for Access Control, to express access policies associated with data stored in decentralised datastores.

4.2.1 Profile requirements specification

This Section outlines the motivation and identified requirements for the development of the OAC profile. As previously mentioned, personal datastores, such as Solid Pods, need to deal with GDPR's requirements, particularly the information requirements set out in Articles 13 and 14, such as the identity of the controller, the purpose for processing, the personal data categories being processed, or the legal basis being used, if they are to be adopted as a legally compatible solution for the sharing of personal data in Europe. Taking Solid as a use case, this information can be given to Solid users by employing conventional methods, such as a notice provided through the data requester's website. However, for individuals to control their data practices, the Solid Pod must also record this information so that the individual has the opportunity to:

(i) inspect their personal data within an environment under their control;

(ii) store it for accountability purposes;

(iii) determine their data access preferences; and

(iv) be assisted in enforcing said preferences.

To achieve this, it is necessary to understand the provisions of the law regarding the information that needs to be provided, including the particular requirements of certain legal bases such as

consent, and the forms of control that individuals want to have or the information they want to know in the context of the handling of their personal data. These requirements are based not just on the GDPR, but also on the guidelines of data protection supervisory authorities. Therefore, based on these considerations and also on the Solid-related challenges identified in Section 2.4, the usage of ODRL and DPV is motivated by the following needs:

1. Organisations need to:
    a) Specify machine-readable data handling policies, which should be accessible by users;
    b) Document provenance information related to their personal data processing activities, including notices and activity logs;
    c) Determine and fulfil applicable rights and obligations based on specific data protection laws or other contextual information, e.g., specific categories of personal data;
    d) Implement security measures by default and by design, specifically related to personal data access.
2. Users need to:
    a) Express human-centric data-sharing preferences, e.g., willingness to share a specific data type for non-profit research or to prohibit processing for profiling purposes;
    b) Specify broad permissions, e.g., allow data access for scientific research, or restrict third party data collection;
    c) Specify narrow permissions, e.g., allow access to phone contact details for a particular app, or deny access to a specific resource;
    d) Have a policy conflict strategy, e.g., generally deny access to location data, but include an exception for specific applications;
    e) Understand who is using which data categories, for what purposes, sharing it with whom, and under what legal basis.

Moreover, taking into consideration the previously described motivation points, the following requirements can then be specified for the OAC profile:

R1. Support specifying user preferences as policies.

R2. Incorporate vocabulary specifying or aligned to legal concepts.

R3. Support specifying permissions and prohibitions at arbitrary granularity.

R4. Support identifying and resolving conflicts based on scope.

R5. Record policies used to authorise access to data.

R6. Support querying policies and authorisations for introspection of data access.

As such, following the LOT methodology, these requirements are consolidated in the profile's ORSD available in Table 4.4. As Solid's current access control mechanisms only partially implement R1, R3, and R5, OAC allows its users to declare not only granular permissive policies but also prohibitive policies, both aligned with legal requirements, which can be stored in their decentralised datastores

for future inspection and can be used with additional constraints and contextual information. Moreover, requirement R1 is covered by the competency questions CQO1 to CQO5, R2 by CQO8 and CQO9, R4 by CQO6, and R3, R5, and R6 by all CQs in Table 4.4.

Table 4.4: Ontology Requirement Specification Document of the OAC profile.

ODRL Profile for Access Control	
1. Purpose	
The purpose of this profile of ODRL is to support policies determining the access to personal data stored in decentralised storage environments, such as Solid Pods.	
2. Scope	
The scope of this profile is limited to the definition of an ODRL Profile for Access Control in decentralised settings. In particular, the introduced elements will serve one of these purposes: (i) define actions supporting the enforcement of current ACL verbs, (ii) define data protection-related actions and restrictions defined in GDPR, (iii) any vocabulary element to support policy patterns that can be anticipated to be common, and (iv) elements necessary to support the authorisation reasoning decision.	
3. Implementation Language	
RDF, RDFS	
4. Intended End-Users	
Developers of decentralised storage servers and applications, such as Solid servers and apps.	
5. Intended Uses	
Use 1. Declaration of a policy by an individual storing personal data in a decentralised datastore, such as a Solid Pod. Use 2. Request of data made by an entity, service, or application to gain access to the data in different modalities. Use 3. Records of data access with transparent information related to the policy matching algorithm, including contextual information.	
6. Ontology Requirements	
a. Non-Functional Requirements	
NFR 1. The profile is published online with HTML documentation, following W3C's specification format.	
b. Functional Requirements: Groups of Competency Questions	
CQOG1. Related to access	CQOG2. Related to GDPR
CQO1. Which policy type is being defined? CQO2. Which actions are defined in the policy? CQO3. Which data types are mentioned in the policy? CQO4. Which policy constraints need to be fulfilled? CQO5. Who are the parties intervening in the policy? CQO6. Which is the conflict strategy of a policy? CQO7. What are the contextual elements that need to be considered in the policy matching algorithm?	CQO8. Which information about personal data and its processing is necessary to have legally aligned policies? CQO9. What identification information needs to be provided by the policy parties?

Lastly, it should be clear that OAC is **not dependent on Solid**, as it is not based on any Solid-specific vocabularies, and can be used in other decentralised data environments. This platform independence is an important component of OAC as OAC can be used in any personal datastores that store primarily personal data, with the goal of enabling data subjects to understand and control policies to exercise their individual GDPR rights and respect GDPR principles. Nevertheless, throughout this Thesis, the use of OAC is demonstrated through the Solid ecosystem as it is an example of the implementation of a decentralised environment for the sharing of (personal) data that is based on the Semantic Web stack of technologies.

4.2.2 Profile implementation

This ODRL profile relies on DPV, for the invocation of legal concepts related to data protection and privacy, and ACL, for the expression of access mode operations, to specify complex permissions, prohibitions, or duties over the access to personal data resources.

Moreover, OAC policies can be used to add a new layer to decentralised data systems – a layer that is currently missing from the Solid ecosystem for instance – that will come between the data and the access authorisation, e.g., ACL or ACP authorisations, layers in order to provide a richer access control mechanism to such systems. As an access control mechanism's main goal is to determine access by users or software agents to digital resources, the entities generating and/or providing the data must able to express policies that satisfy their preferences, while users or software agents who wish to access said data must be able to define policies that describe their data handling activities. By using these policies in an algorithm to match incoming access requests for data, an agreement over the access to a certain resource or type of data can be defined and the decentralised data system can provide a fine-grained access control mechanism to its users. As such, OAC reuses ODRL's `Offer` policies to express the conditions for access to personal data stored in decentralised data systems, e.g. Solid Pods, `Request` policies to express users or software agents' access requests and `Agreement` policies to describe the agreed conditions for access to the data. The three types of policies are defined below, according to their definition provided in the ODRL Vocabulary & Expression 2.2 Recommendation specification [Iannella et al., 2018]:

- **Offer** – Policy that proposes the assigner's rules over an asset and does not grant any privileges to assignees.
- **Request** – Policy that proposes the assignee's rules over an asset and does not grant any privileges to any parties.
- **Agreement** – Policy issued by an assigned that grants privileges to the assignee over an asset.

OAC's core concepts are illustrated in Figure 4.3 and Tables 4.5 and 4.6 specify the alignment between the ODRL, DPV, and ACL terms to ensure that their semantics are correctly interpreted by OAC implementers.

Two new types of policies, which can be combined in ODRL offers, are specified to deal with the preferences and requirements of users who wish to define rules for the processing of their personal data:

- **Preference** – Soft policy that expresses the assigner's preferences over a personal data asset which may not be satisfied and must not grant any privileges to assignees. If a preference policy set by party A does not match a request policy from party B, the request can still be accepted if party A accepts party B's request conditions.
- **Requirement** – Hard policy that expresses the assigner's preferences over a personal data asset which must be satisfied and must not grant any privileges to assignees. If a requirement policy set by party A does not match a request policy from party B, the request must be denied even if party A accepts party B's request conditions.

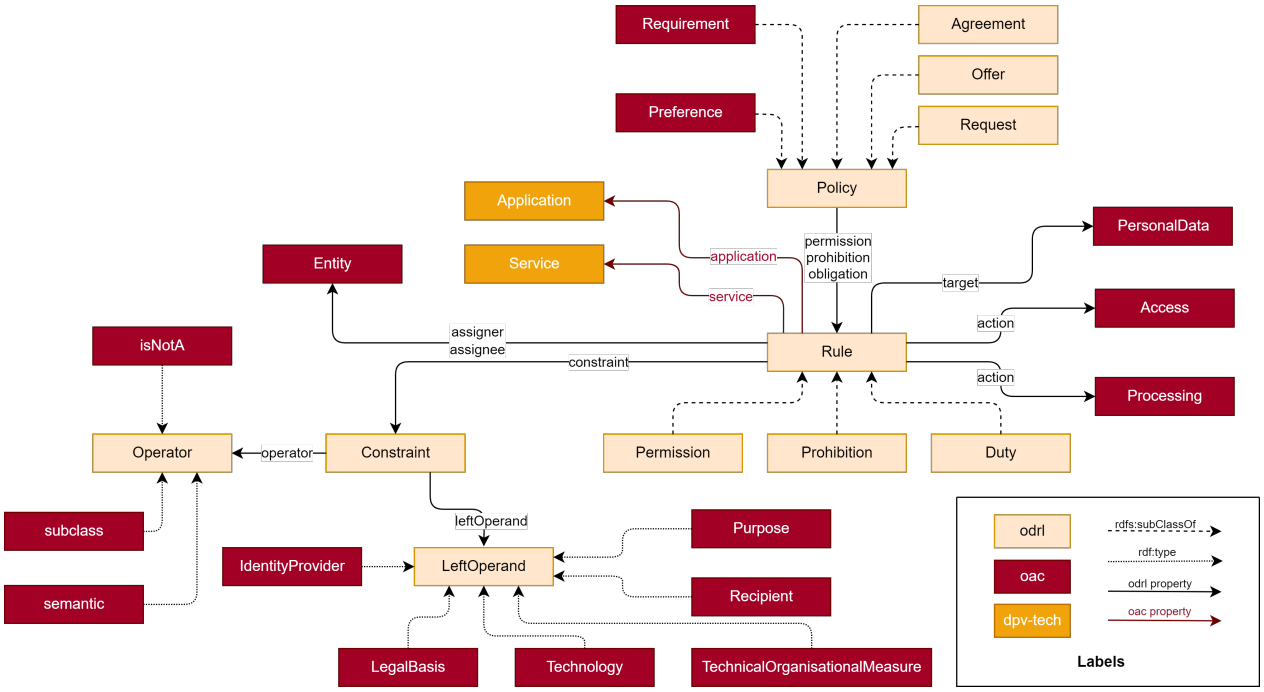


Figure 4.3: Diagrams of the concepts specified by the OAC profile.

Table 4.5: Classes and named individuals specified in the OAC profile.

Profile term	Instance of	Subclass of
oac:Preference		odrl:Policy
oac:Requirement		odrl:Policy
oac:isNotA	odrl:Operator	
oac:subclass	odrl:Operator	
oac:semantic	odrl:Operator	
oac:PersonalData	odrl:Asset	dpv:PersonalData
oac:Access	odrl:Action	acl:Access
oac:Processing	odrl:Action	dpv:Processing
oac:Entity	odrl:Party	dpv:Entity
oac:Purpose	odrl:LeftOperand	dpv:Purpose
oac:Recipient	odrl:LeftOperand	dpv:Recipient
oac:LegalBasis	odrl:LeftOperand	dpv:LegalBasis
oac:TechnicalOrganisationalMeasure	odrl:LeftOperand	dpv:TechnicalOrganisationalMeasure
oac:Technology	odrl:LeftOperand	dpv:Technology
oac:IdentityProvider	odrl:LeftOperand	

Table 4.6: Properties specified in the OAC profile.

Profile property	Domain	Range
oac:service	odrl:Rule, odrl:Policy	dpv-tech:Service
oac:application	odrl:Rule, odrl:Policy	dpv-tech:Application

Listing 4.1 presents an example of an OAC requirement and an OAC preference policies and Listing 4.2 an ODRL offer, based on the previously listed requirement and preference policies, as is indicated by the `dcterms:source` property. The permission associated with the requirement policy contains the property `dpv:hasContext` associated with the term `dpv:Required` to indicate that said permission is a requirement, while the term `dpv:Optional` is used to identify the rules related with a preference policy.

Listing 4.1 OAC requirement and preference policies issued by https://solidweb.me/besteves4/profile/card#me.

```
<https://solidweb.me/besteves4/policies/requirement1> a oac:Requirement ;
    odrl:uid <https://solidweb.me/besteves4/policies/requirement1> ;
    odrl:profile oac: ;
    dcterms:description "Requirement to read identifier data for identity
     ↪ verification purposes." ;
    dcterms:creator <https://solidweb.me/besteves4/profile/card#me> ;
    dcterms:issued "2023-10-20T18:22:15"^^xsd:dateTime ;
    odrl:permission [
        odrl:assigner <https://solidweb.me/besteves4/profile/card#me> ;
        odrl:target oac:Identifier ;
        odrl:action oac:Read ;
        odrl:constraint <#Constraint_Purpose_IdentityVerification> .

<#Constraint_Purpose_IdentityVerification> a odrl:Constraint ;
    dcterms:title "Purpose for access is to verify the identity of the
     ↪ assigner." ;
    odrl:leftOperand oac:Purpose ;
    odrl:operator odrl:isA ;
    odrl:rightOperand dpv:IdentityVerification .

<https://solidweb.me/besteves4/policies/preference1> a oac:Preference ;
    odrl:uid <https://solidweb.me/besteves4/policies/preference1> ;
    odrl:profile oac: ;
    dcterms:description "Preference to read age data if purpose is not
     ↪ commercial research." ;
    dcterms:creator <https://solidweb.me/besteves4/profile/card#me> ;
    dcterms:issued "2023-10-20T18:26:09"^^xsd:dateTime ;
    odrl:permission [
        odrl:assigner <https://solidweb.me/besteves4/profile/card#me> ;
        odrl:target oac:Age ;
        odrl:action oac:Read ;
        odrl:constraint <#Constraint_Purpose_not_CommercialResearch> .

<#Constraint_Purpose_not_CommercialResearch> a odrl:Constraint ;
    dcterms:title "Purpose for access is not commercial research." ;
    odrl:leftOperand oac:Purpose ;
    odrl:operator oac:isNotA ;
    odrl:rightOperand dpv:CommercialResearch .
```

Additionally, a set of three new ODRL operators, which are currently missing from the ODRL Core vocabulary Recommendation, and two new properties to specify policies applicable to certain services or applications, `oac:service` and `oac:application`, which are important stake-

Listing 4.2 ODRL offer issued by https://solidweb.me/besteves4/profile/card#me.

```
<https://solidweb.me/besteves4/policies/offer1> a odrl:Offer ;
    odrl:uid <https://solidweb.me/besteves4/policies/offer1> ;
    odrl:profile oac: ;
    dcterms:description "Offer to read identifier data for identity
     ↪ verification and age data if purpose is not commercial research."
     ↪ ;
    dcterms:creator <https://solidweb.me/besteves4/profile/card#me> ;
    dcterms:source <https://solidweb.me/besteves4/policies/requirement1>,
     ↪ <https://solidweb.me/besteves4/policies/preference1> ;
    dcterms:issued "2023-10-20T22:15:34"^^xsd:dateTime ;
    odrl:permission [
        dpv:hasContext dpv:Required ;
        odrl:assigner <https://solidweb.me/besteves4/profile/card#me> ;
        odrl:action oac:Read ;
        odrl:target oac:Identifier ;
        odrl:constraint <#Constraint_Purpose_IdentityVerification>
    ] ;
    odrl:permission [
        dpv:hasContext dpv:Optional ;
        odrl:assigner <https://solidweb.me/besteves4/profile/card#me> ;
        odrl:action oac:Read ;
        odrl:target oac:Age ;
        odrl:constraint <#Constraint_Purpose_not_CommercialResearch>
    ] .

<#Constraint_Purpose_IdentityVerification> a odrl:Constraint ;
    dcterms:title "Purpose for access is to verify the identity of the
     ↪ assigner." ;
    odrl:leftOperand oac:Purpose ;
    odrl:operator odrl:isA ;
    odrl:rightOperand dpv:IdentityVerification .

<#Constraint_Purpose_not_CommercialResearch> a odrl:Constraint ;
    dcterms:title "Purpose for access is not commercial research." ;
    odrl:leftOperand oac:Purpose ;
    odrl:operator oac:isNotA ;
    odrl:rightOperand dpv:CommercialResearch .
```

holders in decentralised data systems, are specified in OAC. The newly introduced `oac:isNotA` operator is used in the `<#Constraint_Purpose_not_CommercialResearch>` constraint, in Listing 4.1, to indicate that the purpose for access can not be an instance of the right operand of the constraint, e.g., `dpv:CommercialResearch`. The `oac:subclass` operator can be used to indicate that a given left operand is a subclass of the right operand of the constraint, e.g., the purpose constraint of a rule can be a subclass of DPV's research and development purpose such as academic research, non-commercial research or commercial research, and the `oac:semantic` operator to express that a given left operand is equal to, an instance or a subclass of the right operand of the constraint, e.g., the purpose constraint of a rule can be research and development, an instance of research and development or one of its subclasses such as academic research, non-commercial research or commercial research.

Personal data is defined as an ODRL asset to define personal data-specific access policies, access modes and processing operations are defined as ODRL actions to define policies for specific access modes and/or processing operations which are not covered by ACL's access modes, e.g., `dpv:Transfer` or `dpv:Copy`, and DPV's `Entity` concept is defined as an ODRL party to define entity-specific access policies. Additionally, when defining ODRL requests, the data requesters might use processing concepts, `dpv:Use, dpv:Collect, dpv:Share`, as the permitted/prohibited action of the rule that differ from the existing ACL's access modes, `acl:Read, acl:Write, acl:Append`. As such, a mapping of ACL verbs to DPV processing operations is provided in OAC for such cases where offers and requests need to be matched and include both ACL access modes and DPV processing operations. In this mapping, the `acl:Read` access mode corresponds to `dpv:Use, dpv:Collect` processing operations, and `acl:Write` resembles `dpv:Store, dpv:MakeAvailable`. Furthermore, as previously mentioned, there are operations such as `dpv:Share` or `dpv:Transfer` that do not have a specific corresponding concept in WAC's ACL vocabulary, which require a greater introspection in the integration of legal processing concepts with access control operations. Moreover, purposes, recipients, legal bases, technical and organisational measures, technologies and identity providers are defined as ODRL constraints to define constraint-restricted access policies.

Listing 4.3 presents an example of an ODRL request that uses OAC terms and Listing 4.4 an ODRL agreement which is the result of the matching between the offer defined in Listing 4.2 and the previously mentioned request. In this example, Beatriz, identified by `https://solidweb.me/besteves4/profile/card#me`, and Arya, identified by `https://solidweb.me/arya/profile/card#me`, reach an agreement to allow read access operations over Beatriz's age data for the purpose of academic research in project X. This `odrl:Agreement` is the result of the matching of `https://solidweb.me/besteves4/policies/offer1` and `https://solidweb.me/arya/requests/age_academicResearch`, as indicated by the `dcterms:references` property. The legal basis of the agreement is consent, as is specified in the policy with the `dpv:hasLegalBasis dpv:Consent` terms, and Beatriz and Arya are registered as the data subject and data controller in question, respectively, using the `dpv:hasDataSubject` and `dpv:hasDataController` terms. Policy matching and agreement generation is discussed in Chapter 6.

This Thesis focuses on *Purpose, Personal Data, Processing, Recipients, Legal Bases, Technical and Organisational Measures* and *Technologies* as the minimum 'core concepts' for the OAC profile, and

Listing 4.3 ODRL request issued by https://solidweb.me/arya/profile/card#me.

```
<https://solidweb.me/arya/requests/age_academicResearch> a odrl:Request ;
    odrl:uid <https://solidweb.me/arya/requests/age_academicResearch> ;
    odrl:profile oac: ;
    dcterms:description "Request to read age data for academic research."
     ↪  ;
    dcterms:creator <https://solidweb.me/arya/profile/card#me> ;
    dcterms:issued "2023-10-21T13:47:56"^^xsd:dateTime ;
    odrl:permission [
        odrl:assignee <https://solidweb.me/arya/profile/card#me> ;
        odrl:action oac:Use ;
        odrl:target oac:Age ;
        odrl:constraint <#Constraint_Purpose_AcademicResearch>
    ] .

<#Constraint_Purpose_AcademicResearch> a odrl:Constraint ;
    dcterms:title "Purpose for access is to conduct academic research in
     ↪  project X." ;
    odrl:leftOperand oac:Purpose ;
    odrl:operator odrl:eq ;
    odrl:rightOperand ex:AcademicResearchProjectX .

ex:AcademicResearchProjectX a dpv:Purpose ;
    rdfs:subClassOf dpv:AcademicResearch ;
    rdfs:label "Conduct research in the academic project X." .
```

leaves out other DPV concepts such as rights or risks, which can be added at a later stage if needed. Furthermore, similarly to WAC and ACP, OAC policies can also be defined for particular resources identified by URIs – in such cases when an access request for a particular data type comes in, the authorisation mechanism must have information about what type of data those particular resources contain or else they will not be returned if they match the data type of the request. Such information can be stored in a data registry, stored in a e.g. Solid Pod, where resources can be associated with the type of data they contain by using DPV's `hasPersonalData` property and DPV-PD's taxonomy of personal data categories, e.g., `ex:pod-file1 dpv:hasPersonalData dpv-pd:HealthHistory .`

Ultimately, since these policies are stored in the personal datastore for purposes of accountability and transparency, apps and services, based on the stored preferences, requests, and agreements, can be built, e.g., using SPARQL queries, to inquire who is using what data and for what purposes. Listing 4.5 presents a SPARQL query to retrieve permitted data accesses by user, data, and purpose from ODRL agreements stored in a decentralised datastore.

4.2.3 Profile publication and maintenance

The ontology human-readable documentation and machine-readable file are available at https://w3id.org/oac using content negotiation. The HTML documentation includes a description of the classes and properties of the ontology, that was done in collaboration with domain

Listing 4.4 ODRL agreement to read age data for academic research based on consent.

```
<https://solidweb.me/besteves4/policies/agreement1> a odrl:Agreement ;
    odrl:uid <https://solidweb.me/besteves4/policies/agreement1> ;
    odrl:profile oac: ;
    dcterms:description "Agreement to read age data for academic research
     ↪ based on consent." ;
    dcterms:creator <https://solidweb.me/besteves4/profile/card#me> ;
    dcterms:issued "2023-10-21T13:58:37"^^xsd:dateTime ;
    dcterms:references <https://solidweb.me/besteves4/policies/offer1>,
     ↪ <https://solidweb.me/arya/requests/age_academicResearch> ;
    dpv:hasDataSubject <https://solidweb.me/besteves4/profile/card#me> ;
    dpv:hasDataController <https://solidweb.me/arya/profile/card#me> ;
    dpv:hasLegalBasis dpv:Consent ;
    odrl:permission [
        odrl:assigner <https://solidweb.me/besteves4/profile/card#me> ;
        odrl:assignee <https://solidweb.me/arya/profile/card#me> ;
        odrl:action oac:Read ;
        odrl:target oac:Age ;
        odrl:constraint <#Constraint_Purpose_AcademicResearch>
    ] .
```

Listing 4.5 SPARQL query to retrieve authorised data accesses by user, data, and purpose.

```
SELECT DISTINCT ?User ?Data ?Purpose WHERE {
    ?a a odrl:Agreement .
    ?a odrl:permission ?perm .
    ?perm odrl:assignee ?User .
    ?perm odrl:target ?Data .
    ?perm odrl:constraint ?c .
    ?c odrl:leftOperand oac:Purpose .
    ?c odrl:operator odrl:eq .
    ?c odrl:rightOperand ?Purpose .
}
```

experts[1], a diagram with the graphical representation of the ontology, examples of policies defined with the OAC profile, and information related to the policy matching algorithm. The ontology documentation also includes metadata, such as the identity of the creators and publishers of the ontology, the dates of creation and last modification, or the version number.

The source code is hosted at https://w3id.org/oac/repo, under the CC-BY-4.0 license. The repository can also be used by OAC users to suggest new inclusions to the ontology and to report bugs through GitHub Issues. In addition, the repository at https://w3id.org/oac/policies contains a growing collection of OAC policies that can be reused by OAC users.

[1]This collaboration was performed with legal and ontology engineering experts in the context of W3C's ODRL and DPV community groups.

4.3 Metadata language for Solid

This Section describes the development of PLASMA, a metadata language for policy-based access control in Solid, to express metadata related to the entities, registries, logs, policies and infrastructure necessary to provide transparency to Solid's data handling practices.

4.3.1 PLASMA requirements specification

This Section outlines the motivation and identified requirements for the development of PLASMA, a Policy LAnguage for Solid's Metadata-based Access control. As previously mentioned, Solid builds upon Web's ethical principles[2] and standards such as LDP or RDF and, in accordance with its Protocol, relies on said standards to *"realise a space where individuals can maintain their autonomy, control their data and privacy, and choose applications and services to fulfil their needs"* [Capadisli et al., 2022].

Although it was designed with these goals in mind, Solid currently lacks compatibility with data protection regulatory efforts [Pandit, 2023], such as the GDPR. In particular, Solid lacks a practical mechanism to enforce GDPR's principles of transparency and accountability as there are no tools for users, applications, or services to model or document information related to privacy notices, agreements, consent and rights exercising. Furthermore, Solid is based on a ground-up redesign where machine-readable information is encouraged to be provided and reused towards improving the value of data and quality of life for users. However, Solid's access control specifications do not contain any mechanism by which apps can provide or users can understand or express information regarding who/why/how data will be used, and to utilise these in making the process of granting and controlling access to data easier and legally compatible. This lack of 'actionable records' also strengthens the propagation of existing problems of the Web such as the use of dark patterns or manipulations to gain access to personal data of Web users.

Given that users are well versed in the usage of apps, e.g. on their smartphones, there is an expectation that Solid should also adopt an environment of trust and accountability that reduces the cognitive overload on users to understand complex information and make informed decisions, and where the environment guides responsible and accountable development. Examples of such measures include the usage of app stores and curated or approved application verification processes. Without these, Solid users currently have no means to identify who are the actors behind the app and authorities cannot know whom to approach when opening an investigation on faulty data practices.

Therefore, based on these considerations, the following requirements were drafted for the development of PLASMA:

R1. Support specifying information about Solid infrastructure.

R2. Record information about Pod, apps, services, and data providers/developers.

R3. Support specifying of different agreements and notices.

R4. Record provenance information for future introspection and convenient access to data.

[2] https://www.w3.org/TR/ethical-web-principles/ (accessed on 22 October 2023)

R5. Provide conformance conditions to assist with legal compliance.

As such, following the LOT methodology, these requirements are consolidated in the ORSD available in Table 4.7. By incorporating the usage of PLASMA, Solid actors can describe their data practices in a responsible and accountable manner in a way that addresses the above-mentioned requirements. As such, requirement R1 is covered by the competency questions CQP1 and CQP5, R2 by CQP2 to CQP6, R3 by CQP3 and CQP4, R4 by CQP1 to CQP8, and R5 by CQP9. In addition to providing the vocabulary, PLASMA also demonstrates how a *decentralised ecosystem* can be developed that takes advantage of the machine-readable nature of RDF information, such as to guarantee that apps declare a set of metadata before being allowed access to data, and ensures that apps, services, agents, Pods, and users act in conformance and provide an environment of trust and accountability.

Table 4.7: Ontology Requirement Specification Document of PLASMA.

Policy LAnguage for Solid's Metadata-based Access control
1. Purpose
The purpose of PLASMA is to provide consistent taxonomies to describe the entities, infrastructure, policies, notices, registries and logs necessary to understand and establish responsibilities and accountability within the Solid ecosystem.
2. Scope
The scope of this ontology is limited to the definition of a metadata language to provide transparency Solid's data handling practices. PLASMA promotes the usage of OAC to determine access control to Solid Pod's resources, provides conformance conditions and workflow scenarios where PLASMA terms should be used.
3. Implementation Language
RDF, RDFS
4. Intended End-Users
Developers of Solid servers, applications, services or agents.
5. Intended Uses
Use 1. Describing entities, infrastructure and processes involved in the Solid ecosystem. Use 2. Expressing information regarding legal roles and other compliance requirements in a jurisdiction-agnostic manner (while satisfying requirements from GDPR). Use 3. Defining patterns for the expression of users and apps policies, data use logs, and registries to provide easy access to data in Pods.
6. Ontology Requirements
a. Non-Functional Requirements
NFR 1. The ontology is published online with HTML documentation, following W3C's specification format.
b. Functional Requirements: Groups of Competency Questions
CQP1. Which Pod management data is stored in the Pod? CQP2. Which metadata should be recorded when data is added/updated/removed to/from the Pod? CQP3. What data, including policies, are available in the Pod? CQP4. What policy describes the data access requirements of a certain app or service? CQP5. Who are the parties providing Pod infrastructure? CQP6. How and where is the data being physically stored? CQP7. What registries are available in the Pod for convenient access to data? CQP8. What identification information needs to be provided by Solid-involved parties? CQP9. Which information about personal data processing is necessary to have legally aligned decentralised datastores?

4.3.2 PLASMA taxonomies

PLASMA relies on OAC for the expression of policies related to access to personal data stored in Solid Pods, on the W3C Recommendation DCAT (Data CATalog vocabulary) [Albertoni et al., 2020] for the expression of data registries and related data sets, on DCMI Metadata Terms [DCMI Usage Board, 2020] for the specification of authorship, temporal and other types of provenance metadata, and on the W3C Recommendation Activity Streams 2.0 [Snell and Prodromou, 2017] for logging relevant events associated with Solid processes. These design choices are aligned with the best practices described in the Data on the Web Best Practices document [Lóscio et al., 2017]. Moreover, while there are legal vocabularies focusing on personal data protection, such as DPV and DPV-GDPR [Pandit et al., 2019b], these are not directly applicable to the Solid ecosystem as the terminology used is not the same, e.g., in Solid, the entity the data belongs/refers to is called *'owner'*, whereas the equivalent term under the GDPR is *'data subject'*. As such, PLASMA provides additional taxonomies to describe the actors, artifacts, and processes involved in the usage of Solid Pods, apps, and services, which are not modelled in the previously mentioned vocabularies. Later on, these can be used to align with their equivalent legal terms (from the GDPR).

Thus, PLASMA supports the implementation of a new *'policy layer'* which aids users, apps, and Pod infrastructure providers to express relevant information about their activities in the form of machine-readable policies, logs, and registers of data. Such sources of information can then be used to enable the development of dashboards for user policy management, machine-readable notifications regarding changes in policies or data, usage of agents to automate tasks, or other interfaces to understand what is being done with the Pod's data. The European Data Protection Supervisor [2021], in its PIMS technical report, mentions these solutions as tools that *"enable individuals themselves to manage and control their online identity"*, by supporting data subjects with consent management, having transparency and traceability to follow their data and their processing, exercising their rights of access, to rectification and erasure, having proof of origin and validity of their data coming from other sources that are not themselves, and providing machine-readability and interoperability in case they want to change storage or other services providers.

Figure 4.4 illustrates the core entities and infrastructure of the Solid ecosystem as specified in PLASMA. Thus, the base PLASMA concepts are defined below as:

- **App** – An application that stores, collects, uses, shares, erases, or performs other actions on Data with the aim of providing specific purposes, services, or functionalities. An application can use several Services and requires human intervention.
- **Service** – A functionality that may or may not utilise or interact with Data within a Pod. Services represent an abstraction of functionality that does not necessarily have to be packaged as an App.

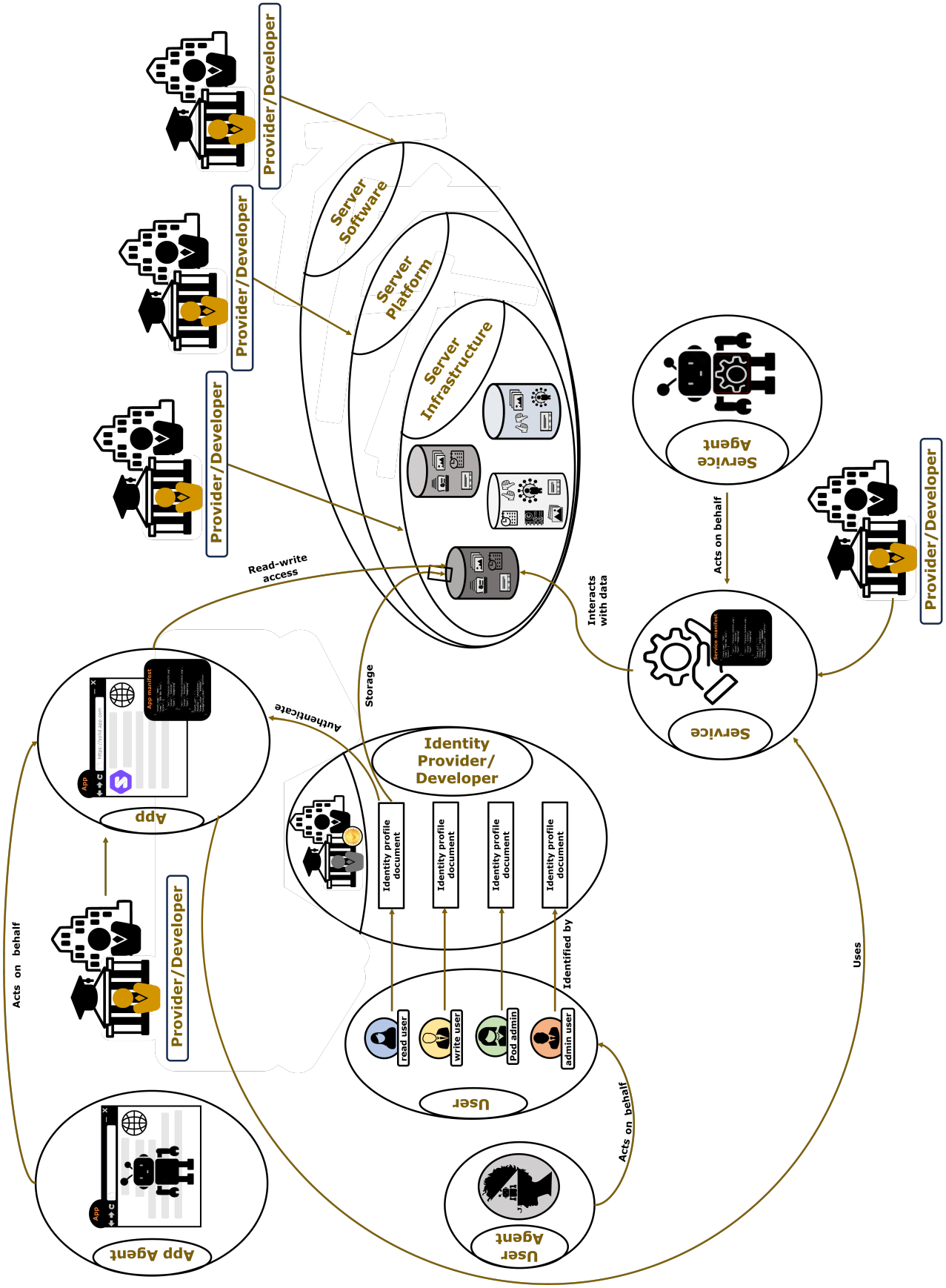


Figure 4.4: Core entities and infrastructure of the Solid ecosystem specified in PLASMA.

- **Pod** – A Personal Data Store that conforms to the Solid Specification.
- **Agent** – A virtual entity associated with carrying out actions within or related to a Pod or its Data.
- **Policy** – A set of guidelines or decisions or recommendations governing the use of Pod or its Data.
- **Data** – Data stored on a Pod or associated with a User, App, Service, or Data Subject of a Pod.
- **Entity** – A legally recognised entity. Legally recognised means the entity has some recognition as being able to enter into agreements, has an address for accountability, and is responsible for obligations and/or rights. Entities are associated with Apps, Services, or Pods.
- **Solid Platform** – The specific implementation of Solid that is installed or used within a Pod.
- **Solid Specification** – The specification that the Pod conforms to in terms of defining the terms, behaviour, and implementation details regarding Pods and their association with Services and Apps.

These terms result from a thorough review of the existing Solid technical documents, in particular of the specifications related to the authorisation protocol, described in Section 2.1.1. While they are mentioned in these specifications, only a slim fraction of them are actually defined in machine-readable form, e.g., ACP provides a `Policy` term and SAI provides a definition for `Application`. Since no concepts were found in the existing Solid vocabularies to define what Pods, services, entities or data, PLASMA provides these terms and extends them with additional taxonomies to cover a wide set of use cases.

Therefore, in the remainder of this Section, an overview of the taxonomies of entities, policies and notices, services, and data defined in the PLASMA vocabulary is provided.

Entities Beyond the agent, client application and identity issuer predicates defined in WAC and/or ACP, there are no other terms to describe the entities involved in the Solid ecosystem. As such, terms to describe the providers and/or developers of Solid apps, services, and other Solid-related processes and infrastructure are provided in PLASMA, as well as terms to describe different types of users and agents. Distinct concepts for providers and developers are specified to distinguish between entities providing the infrastructure, Pods, identity, apps or services, and entities developing them (in case the distinction needs to be made), e.g., when directly using a service from the developer's website, the developer is the same as the provider, however, if a service is being used through a service store or a common marketplace, then the developer is different from the provider and both concepts should be clarified in Solid for a proper allocation of responsibilities.

Furthermore, regarding Solid-related users, an `AdminUser` is defined as a *"User of a Pod that has the administrative capability to make decisions about Data on a Pod"*, i.e., can define who has access to all or parts of Data stored on a Pod, and a `PodAdmin` as a *"User of a Pod that has the*

administrative capability to make decisions about the Pod (separate from Data in a Pod), such as deleting the Pod, changing identity or other resource providers", to effectively distinguish between users who fully control data and users who fully control Pods. This feature is currently missing from the Solid protocol. With regards to the specification of agents, PLASMA departs from Solid's current definition of agent as presently, according to Solid's specifications, they can either be real or virtual agents, e.g., parents on behalf of children or software agents. In PLASMA, agents are *virtual agents*, that can act on behalf of users, apps or services, to distinguish them from entities, which can be held legally responsible.

Figure 4.5 illustrates the providers, developers, users and agents defined in PLASMA.

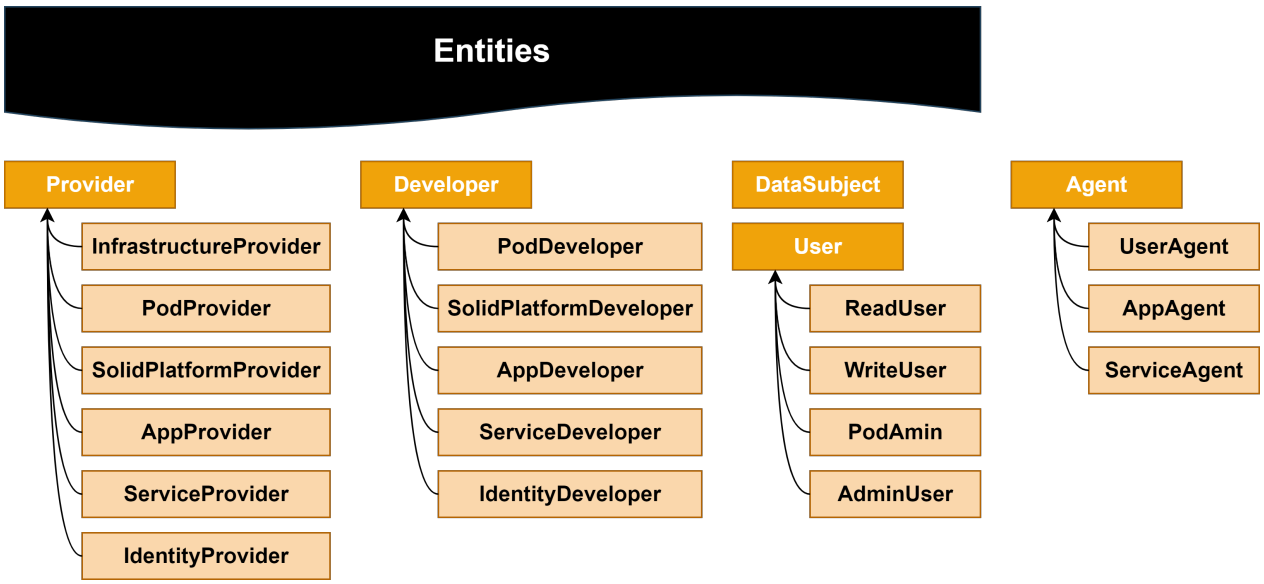


Figure 4.5: Entities and agents specified in PLASMA.

Policies and notices In PLASMA, a policy is a document that specifies user, application, and service requirements for data handling practices that apply to data stored or shared through Solid Pods. Thus, this definition is not limited to access to data stored on a Pod, it also relates to used or transferred data, i.e., to avoid cases such as data collected for purpose X being used for purpose Y. Such a definition aids with the alignment with legal requirements, e.g., it requires that all purposes must be stated at the time of data collection and that requesters need new consent from the data subject if the purpose for access changes. PLASMA provides definitions for user policies, in particular for user offers, requirements and preferences, aligned with the `odrl:Offer`, `oac:Requirement` and `oac:Preference` concepts described in Section 4.2. Regarding data requests, i.e., conditions for apps and services to have access to or use Pod data, PLASMA envisions their integration into the ecosystem by declaring them in a manifest such as the one being conceived in the W3C Web Application Manifest specification – an application manifest is a *"JSON document that contains startup parameters and application defaults for when a web application is launched"* [Cáceres et al., 2023]. However, the current specification is not enough to achieve legal compliance as it does not include information on entities developing the app, their identity and contact details, or their privacy policies. Thus, PLASMA includes app manifest and service

manifest concepts. In addition to user policies and data requests, PLASMA also provides different types of data agreements, a concept that is completely missing from the Solid specifications as they only provide concepts to refer to apps and users' access authorisations. As such, an agreement can either be based on the user's consent, or governed by a contract between the user and the entity having responsibility for the app, service, or Pod.

Additionally, current Solid specifications also lack the definition of notices. Notices are documents that provide context information about entities, operations, or data involved in specific processes, e.g., notices may specify the providers, developers, and/or data handling practices of applications and services. In this regard, PLASMA provides terms for declaring *ex-ante* and *ex-post* notices, i.e., notices declared before and after data access, as well as privacy notices for Pods, users, applications, and services.

Figure 4.6 illustrates the policies and notices concepts specified in PLASMA.

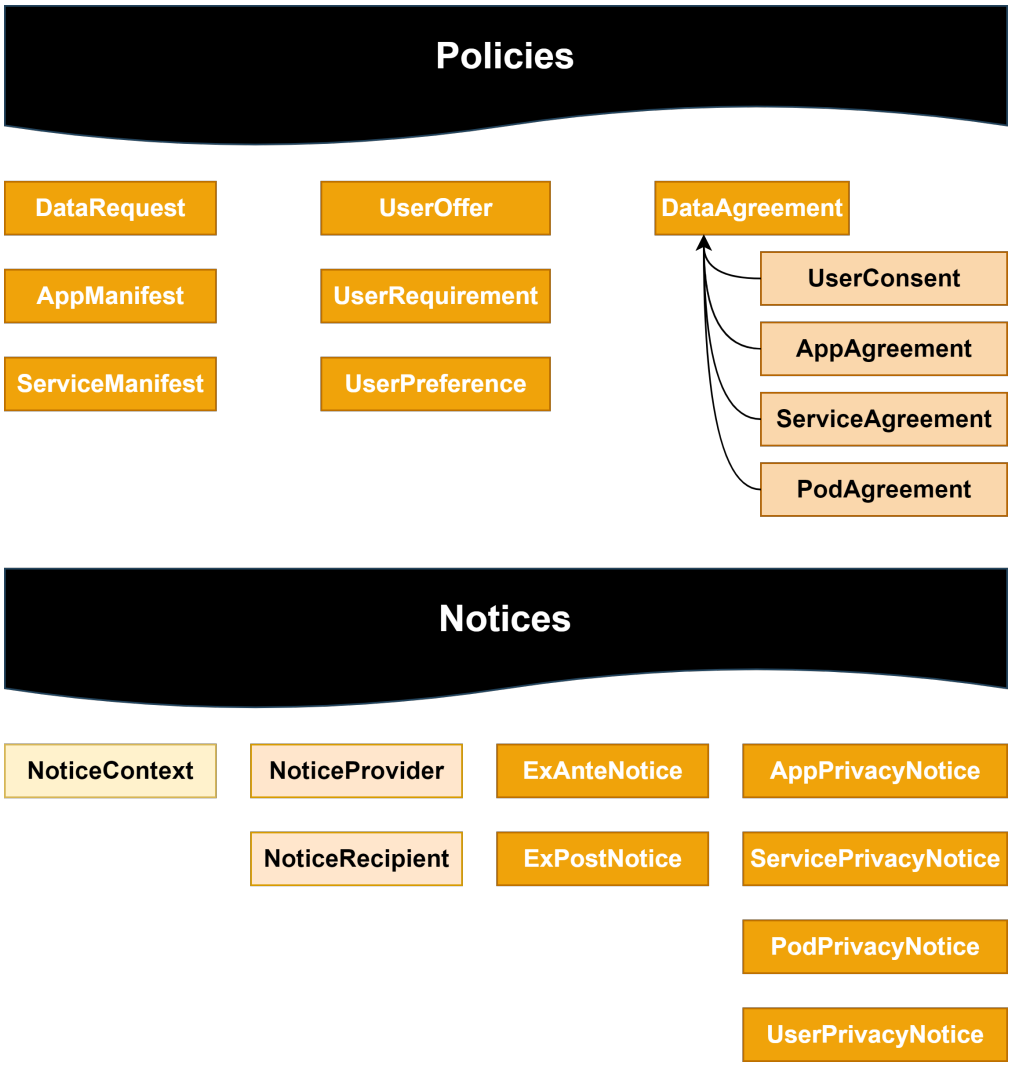


Figure 4.6: Policy types and notices specified in PLASMA.

Services PLASMA differentiates between an app and a service – services convey functionalities that do not need to be packaged as an app. Moreover, an app requires human intervention to perform some action on data for a specific purpose, while a service may not require human intervention to use or interact with data within a Pod. Since services are a new concept being introduced by PLASMA in the Solid ecosystem, a taxonomy of twelveu Solid-related services, which can be further expanded, is supplied and illustrated in Figure 4.7.

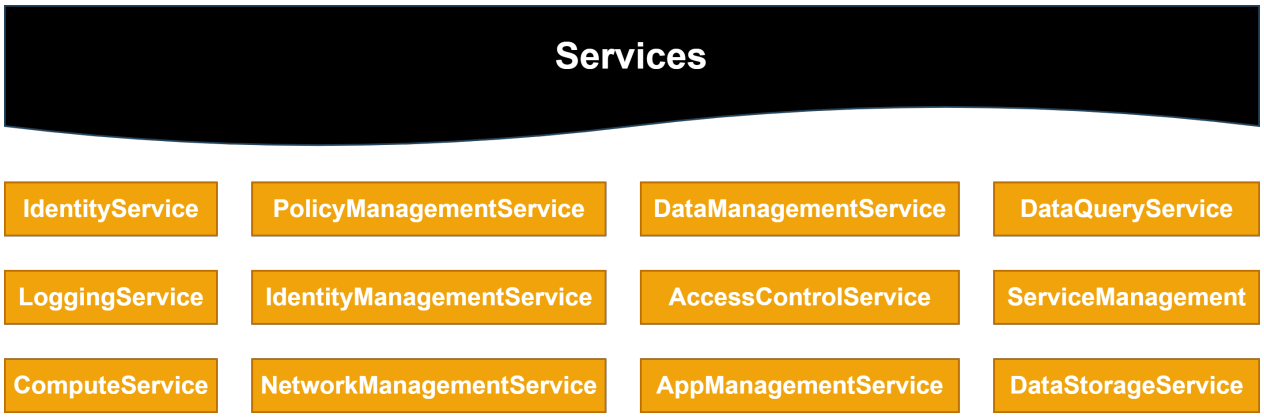


Figure 4.7: Services specified in PLASMA.

Pod-related data To fulfil Solid's vision of providing individuals with a decentralised data storage service for their data and the choice of which applications or services to use for a specific task, metadata regarding Pod management, entities, apps, services, logs and registries should be provided. Additionally, supervisory authorities can use such logging and provenance metadata, from the Pods of users, for auditing activities, e.g., an EU data protection authority can use these records to investigate a personal data breach. To this end, PLASMA includes a collection of Solid-related log terms to record provenance information related to processes such as adding or updating resources in a Pod, i.e., a `DataLog`, registering a policy negotiation procedure reliant on user consent, i.e., a `PolicyLog`, or recording a successful user login operation, i.e., a `IdentityLog`. Furthermore, the maintenance of registries as indexed records for providing collective and convenient access to data within a Pod is of the utmost importance for users, apps, and services to have knowledge of the availability of data categories, supported schemas for data, apps, services, relevant policies, and users that have/had access to data stored within a Pod.

Figure 4.8 illustrates the Pod-related data concepts defined in PLASMA, including logs and registries.

4.3.3 Conformance with PLASMA

In this Section, the conditions for Pods, apps, services, users, and agents to be conformant with PLASMA are described, including information regarding what is mandatory (indicated below by the use of the word *must*) or optional (indicated below by the use of the word *may*) to be provided, and how conformity should be evaluated and assured by implementers of the PLASMA specification. Conformance with these conditions can be checked, e.g., using SHACL shapes.

The W3C Recommendation on Data on the Web Best Practices, which is aimed at the *"publication and usage of data on the Web designed to help support a self-sustaining ecosystem"* [Lóscio et al.,

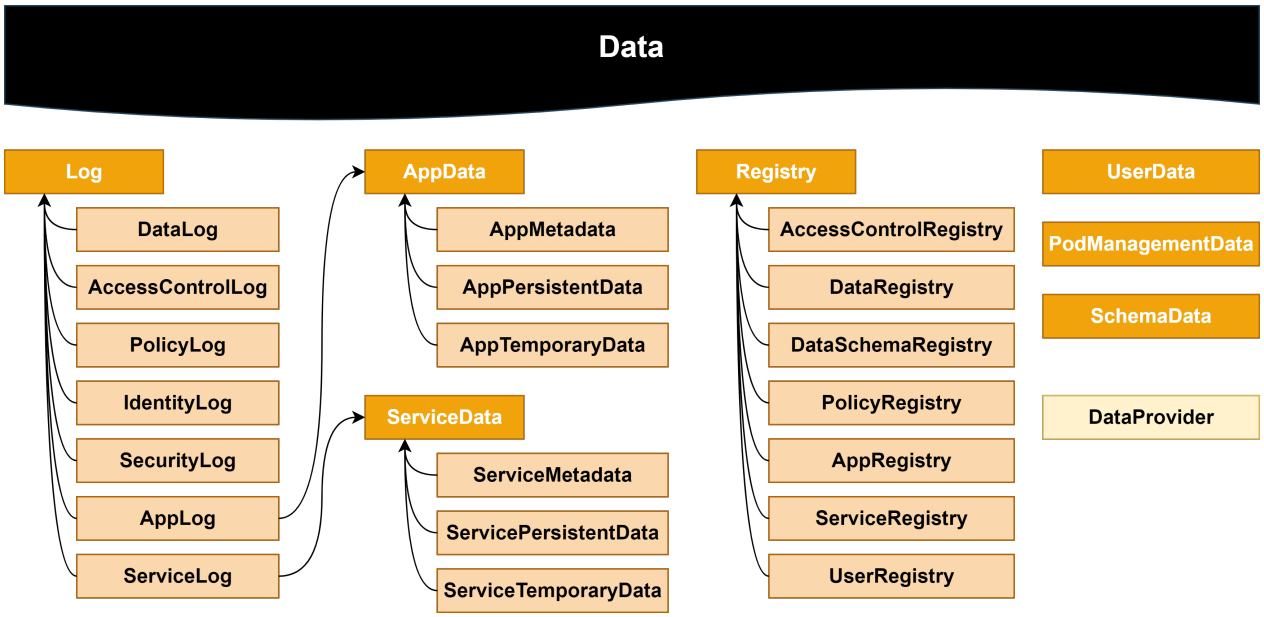


Figure 4.8: Data concepts, including logs and registries, specified in PLASMA.

2017], was followed for the publication of metadata related to provenance, licensing, and versioning of data. The vocabulary specifications for particular tasks, recommended by this best practices document, are mentioned in the following paragraphs. By reusing these concepts, the ontology engineering methodological imperative to maximise the reuse of existing ontological standards to improve interoperability is being followed, as described in the LOT methodology [Poveda-Villalón et al., 2022] followed throughout this Thesis. Table 4.8 summarises the reused vocabularies and respective reused terms. These choices were motivated by the previously mentioned W3C Recommendation on Data on the Web Best Practices, as well as by following ODRL's and DCAT's metadata guidelines, as these are W3C Recommendations for policy expression and dataset, service, and catalog description, respectively. Their usage is further detailed in each of the conformance sections described below.

Table 4.8: Vocabularies reused in PLASMA to express conformity of Pods, apps, services, users, and agents.

Vocabulary	Reused terms
ODRL	`hasPolicy`
DCMI	`description`, `source`, `license`, `conformsTo`, `creator`, `created`, `modified`, `publisher`, `format`, `issued`, `title`, `language`, `type`, `valid`
DPV	`hasName`, `hasContact`, `hasAddress`, `hasPolicy`, `hasNotice`, `hasDataSubject`
Schema.org	`codeRepository`
PAV	`version`
FOAF	`page`
Activity Streams	`summary`, `object`, `actor`, `generator`, `Accept`, `Reject`, `Create`, `Update`, `Delete`, `Move`

Pod conformance For a Pod to be conformant with the PLASMA specification, the following conditions should be satisfied:

- A Pod *must* provide or declare `PodManagementData` which includes metadata about the Pod and of its providers and/or developers, as well as the specific `SolidPlatform` and `SolidSpecification` implemented in the Pod and any `PodAgreement` in place. Listing 4.6 provides an example of the declaration of such metadata.
- A Pod *must* implement or provide equivalent functionality to support the different registries and logs specified in PLASMA. Listing 4.7 provides an example of a data schema registry that records data schemas, formats, or shapes recognised or supported by Pods, apps, or services, as indicated by the `plasma:supportedBy` property.
- A Pod *may* have discovery methods for users to make their data publicly available. These methods *should* rely on the data registries and data schema registries mentioned above.
- A Pod *may* have multiple users with varying degrees of control. A record of the different users and their level of access *must* be kept in the `UserRegistry`.

In addition to the ODRL and DCMI Metadata Terms vocabularies, the DPV's `hasName`, `hasContact` and `hasAddress` properties should be used to identify and provide contact details of the Pod, Solid platform and infrastructure providers and developers. *Schema.org* and the Provenance, Authoring and Versioning (PAV) vocabularies, along with the previously mentioned DCMI, are used to describe the authors, sources, version, and code repository URIs of the platform and specification installed within the Pod. *Schema.org* provides an upper vocabulary of terms to describe *"entities, relationships between entities and actions"* related to structured data on the Web [Guha et al., 2015] and PAV is a *"lightweight ontology for tracking Provenance, Authoring and Versioning"* that *"specializes the W3C provenance ontology PROV-O in order to describe authorship, curation and digital creation of online resources"* [Ciccarese et al., 2013].

Apps and services conformance For an app or service to be conformant with the PLASMA specification, the following conditions should be satisfied:

- An app, or service, *must* have an `AppManifest`, or `ServiceManifest`, in conformance with the W3C Web Application Manifest [Cáceres et al., 2023]. Listing 4.8 provides an example of such an app manifest.
- An `AppManifest`, or `ServiceManifest`, *must* include information regarding the developer and provider of legally relevant entities and their identities.
- An `AppManifest`, or `ServiceManifest`, *must* state the `DataRequest` representing the request to use data using the `odrl:hasPolicy` property. The request *must* provide all information regarding the use of data even if only some of it will be applicable initially or used in the notice.
- An `AppManifest`, or `ServiceManifest`, *must* link the privacy notice of the app, or service, using the `dpv:hasNotice` property. The Pod *may* use this information to display or optionally construct its own notice based on the preferences or accessibility requirements of the user.

Listing 4.6 Metadata of Beatriz's Pod.

```
<https://solidweb.me/besteves4/PodMetadata> a plasma:PodManagementData ;
    dcterms:description "Metadata of Beatriz's Pod" ;
    odrl:hasPolicy <https://solidweb.me/besteves4/agreements/Pod> ;
    plasma:hasProvider
     ↪ <https://solidweb.me/besteves4/entities/PodProvider> ;
    plasma:hasProvider
     ↪ <https://solidweb.me/besteves4/entities/PlatformProvider> ;
    plasma:hasProvider
     ↪ <https://solidweb.me/besteves4/entities/InfrastructureProvider> ;
    plasma:implementedSolidPlatform
     ↪ <https://solidweb.me/besteves4/Platform> ;
    plasma:implementedSolidSpecification
     ↪ <https://solidweb.me/besteves4/SolidSpec> .

<https://solidweb.me/besteves4/agreements/Pod> a plasma:PodAgreement .

<https://solidweb.me/besteves4/entities/PodProvider> a plasma:PodProvider
 ↪ ;
    dpv:hasName "Entity A" ; dpv:hasContact "mailto:entity_a@mail.com" ;
    dpv:hasAddress "Address of Entity A" .

<https://solidweb.me/besteves4/entities/PlatformProvider> a
 ↪ plasma:SolidPlatformProvider .

<https://solidweb.me/besteves4/entities/InfrastructureProvider> a
 ↪ plasma:InfrastructureProvider .

<https://solidweb.me/besteves4/Platform> a plasma:SolidPlatform ;
    plasma:hasProvider
     ↪ <https://solidweb.me/besteves4/entities/PlatformProvider> ;
    dcterms:source <https://communitysolidserver.github.io> ;
    dpv:hasPolicy
     ↪ <https://www.serverproject.de/files/solidweb_me_terms.txt> ;
    schema:codeRepository
     ↪ <https://github.com/CommunitySolidServer/CommunitySolidServer> ;
    pav:version "6.1.0" ;
    dcterms:license <https://dalicc.net/licenselibrary/MIT> .

<https://solidweb.me/besteves4/SolidSpec> a plasma:SolidSpecification ;
    dcterms:conformsTo
     ↪ <https://solidproject.org/TR/2022/protocol-20221231> ;
    dcterms:creator "Sarven Capadisli", "Tim Berners-Lee", "Ruben
     ↪ Verborgh", "Kjetil Kjernsmo" ;
    dcterms:license <https://dalicc.net/licenselibrary/MIT> ;
    pav:version "0.10.0" ; dcterms:created "2022-12-31"^^xsd:date ;
    schema:codeRepository <https://github.com/solid/specification> .
```

Listing 4.7 Data schema registry of Beatriz's Pod.

```
<https://solidweb.me/besteves4/SchemaRegistry> a
    plasma:DataSchemaRegistry ;
    dcterms:description "Registry listing recognised or supported
        schemas" ;
    dcterms:created "2023-09-10T11:51:17"^^xsd:dateTime ;
    dcterms:modified "2023-10-07T12:39:50"^^xsd:dateTime ;
    dcterms:publisher
        <https://solidweb.me/besteves4/entities/DataMgtServiceProvider> ;
    plasma:hasSchema <https://solidweb.me/besteves4/schemas/EHR-schema>,
        <https://solidweb.me/besteves4/schemas/img-format>,
        <https://solidweb.me/besteves4/schemas/entity-shape> .

<https://solidweb.me/besteves4/entities/DataMgtServiceProvider> a
    plasma:ServiceProvider ;
    plasma:serviceType plasma:DataManagementService ;
    dpv:hasName "Entity A" ;
    dpv:hasAddress "Address of Entity A" ;
    dpv:hasContact "mailto:entity_a@mail.com" .

<https://solidweb.me/besteves4/schemas/EHR-schema> a plasma:SchemaData ;
    dcterms:conformsTo <http://www.w3.org/TR/turtle/> ;
    plasma:supportedBy <https://example.com/health-service> .

<https://example.com/health-service> a plasma:Service .

<https://solidweb.me/besteves4/schemas/img-format> a plasma:SchemaData ;
    dcterms:format
        <https://www.iana.org/assignments/media-types/image/png>,
        <https://www.iana.org/assignments/media-types/image/svg+xml> ;
    plasma:supportedBy <https://example.com/social-app> .

<https://example.com/social-app> a plasma:App .

<https://solidweb.me/besteves4/schemas/entity-shape> a plasma:SchemaData,
    sh:NodeShape ;
    plasma:supportedBy <https://solidweb.me/besteves4/> ;
    sh:name "PLASMA entity shape" ;
    sh:description "Minimum data that PLASMA entities should provide to
        be identified." ;
    sh:targetClass plasma:Entity .
```

- An `AppManifest`, or `ServiceManifest`, *must* be stored in the Pod `AppRegistry`, or `ServiceRegistry`. Listing 4.9 provides an example of an app registry, where app-related metadata, including the manifest, app providers and developers, temporary or persistent app data, is recorded.
- Apps, or services, *may* have multiple `AppAgents`, or `ServiceAgents`, which *must* be registered in the `AppRegistry`, or `ServiceRegistry`.

In addition to the PLASMA terms mentioned in the previous list, the usage of the `serviceType` property can be used to connect service providers and developers with a particular type of service, e.g., from PLASMA's service taxonomy, that is provided or developed by said entity. Moreover, it should be noted that app or service stores, such as those maintained by Apple and Google, can also act as app or service providers and the FOAF `page` property can be used to actually connect the store provider with the store itself. The FOAF vocabulary specification [Brickley and Miller, 2004] provides terms to describe people and related personal information and online accounts. Regardless, support for other optional properties specified in the W3C Web Application Manifest specification, such as icons, display mode, orientation, and background colour, can also be integrated into the modelled manifests, as well as 'common' app store metadata, such as screenshots, user rating or app type, e.g., health, game or news app [Gustafson, 2023].

Listing 4.8 App manifest of Contacts app.

```
<https://example.com/Contacts> a plasma:App ;
    plasma:hasAppManifest <https://example.com/Contacts/Manifest> .

<https://example.com/Contacts/Manifest> a plasma:AppManifest ;
    dcterms:conformsTo <https://www.w3.org/TR/appmanifest/> ;
    dcterms:issued "2023-10-23T22:43:58"^^xsd:dateTime ;
    dcterms:title "Contacts" ;
    dcterms:description "App to manage contacts" ;
    dcterms:language <http://id.loc.gov/vocabulary/iso639-1/en> ;
    plasma:hasProvider <https://solidweb.me/besteves4/entities/AppStore>
     ↪  ;
    plasma:hasDeveloper
     ↪  <https://solidweb.me/besteves4/entities/ContactsDeveloper> ;
    odrl:hasPolicy <https://example.com/Contacts/Request> ;
    dpv:hasNotice <https://example.com/Contacts/Notice> .

<https://solidweb.me/besteves4/entities/AppStore> a plasma:AppProvider ;
    dpv:hasName "App Store provider" ;
    dpv:hasAddress "Address of App Store provider" ;
    dpv:hasContact "mailto:app_store@mail.com" ;
    foaf:page <https://example.com/AppStore> ;
    dpv:hasNotice <https://example.com/AppStore/PrivacyPolicy> .

<https://solidweb.me/besteves4/entities/ContactsDeveloper> a
 ↪  plasma:AppDeveloper .

<https://example.com/Contacts/request> a plasma:DataRequest, odrl:Request
 ↪  .
```

Listing 4.9 App registry of Beatriz's Pod.

```
<https://solidweb.me/besteves4/AppRegistry> a plasma:AppRegistry ;
    dcterms:description "Registry listing apps" ;
    dcterms:created "2023-09-30T11:33:35"^^xsd:dateTime ;
    dcterms:modified "2023-10-07T11:31:40"^^xsd:dateTime ;
    dcterms:publisher
     ↪ <https://solidweb.me/besteves4/entities/AppMgProvider> ;
    plasma:hasApp <https://solidweb.me/besteves4/apps/Contacts/> .

<https://solidweb.me/besteves4/entities/AppMgProvider> a
 ↪ plasma:ServiceProvider ;
    plasma:serviceType plasma:AppManagementService .

<https://solidweb.me/besteves4/apps/Contacts/> a plasma:App ;
    plasma:hasAppManifest
     ↪ <https://solidweb.me/besteves4/apps/Contacts/Manifest> ;
    odrl:hasPolicy
     ↪ <https://solidweb.me/besteves4/apps/Contacts/Agreement> ;
    plasma:hasAppMetadata
     ↪ <https://solidweb.me/besteves4/apps/Contacts/Metadata> ;
    plasma:hasAppPersistentData
     ↪ <https://solidweb.me/besteves4/apps/Contacts/PersistentData> ;
    plasma:hasAppTemporaryData
     ↪ <https://solidweb.me/besteves4/apps/Contacts/TemporaryData> ;
    plasma:hasAgent <https://solidweb.me/besteves4/apps/Contacts/Agent> .

<https://solidweb.me/besteves4/apps/Contacts/Metadata> a
 ↪ plasma:AppMetadata ;
    dcterms:description "Contacts metadata" ;
    plasma:hasProvider <https://solidweb.me/besteves4/entities/AppStore>
     ↪ ;
    plasma:hasDeveloper
     ↪ <https://solidweb.me/besteves4/entities/ContactsDeveloper> .

<https://solidweb.me/besteves4/apps/Contacts/PersistentData> a
 ↪ plasma:AppPersistentData ;
    dcterms:type pd:TelephoneNumber ; rdf:value "(+34)691485135" .

<https://solidweb.me/besteves4/apps/Contacts/TemporaryData> a
 ↪ plasma:AppTemporaryData ;
    dcterms:type pd:TelephoneNumber ; rdf:value "(+34)691998745" ;
    dcterms:valid "2023-10-01T14:50:21"^^xsd:dateTime .

<https://solidweb.me/besteves4/apps/Contacts/Agent> a plasma:AppAgent .
```

User conformance For a user to be conformant with the PLASMA specification, the following conditions should be satisfied:

- Impactful interactions of a user, e.g., changing identity providers, *must* be recorded using a well-defined shape. Listing 4.10 provides an example of access control logs modelled with PLASMA and using W3C's Activity Streams activity types [Snell and Prodromou, 2017].
- A `UserRegistry` *must* contain information regarding the `DataSubjects` storing data within a Pod, the `PodAdmin` and other users accessing Data. Listing 4.11 provides an example of said registry.
- Users *may* be directly associated with a `DataRequest` so that they can make requests for data without using an application or service.
- Users *may* have multiple `UserAgents`, which should be registered in the `UserRegistry`.

PLASMA recommends the usage of the W3C's Activity Streams vocabulary [Snell and Prodromou, 2017] to model the distinct logs that should be stored in Solid Pods for transparency and accountability, e.g., data, identity, or policy logs. This recommendation provides an extensive list of activity, actor, and object types that provides *"a baseline extensible syntax for the expression of completed activities"*. Such syntax allows the identification of the resource being logged, using the `as:object` property, the entity responsible for the activity being logged, using the `as:actor` property, or the app or service used to generate the log, employing the `as:generator` property. The `as:Accept` and `as:Reject` activity types can be reused to express when access to data or policies is accepted or rejected, and the `as:Create`, `as:Update`, `as:Delete` and `as:Move` activities can be reused to log the creation, modification, deletion or movement of data or policies. New activity types, e.g., to request data, change identity providers, or verify the identity of apps, which are not modelled by the Activity Streams vocabulary, are modelled in PLASMA, e.g., as `plasma:Request`, `plasma:ChangeIdP` or `plasma:Verify`.

It should also be stated that Inrupt's Enterprise Solid Server has started to provide an auditing service[3] which also relies on the W3C Activity Streams 2.0 Recommendation [Snell and Prodromou, 2017] to document audit events related with the identity of user and applications.

Agent conformance For an agent to be conformant with the PLASMA specification, the following conditions should be satisfied:

- Agents activity *must* be in accordance with the manifest of the entity for which they are acting on behalf of. As such, agents *must* follow the policies of the entities they are acting on behalf of.
- A record of the usage of an `AppAgent`, `ServiceAgent` or `UserAgent` *must* be kept in the `AppRegistry`, `ServiceRegistry` or `UserRegistry`, respectively, including information regarding its providers/developers for accountability. User `https://solidweb.me/besteves4/entities/DataSubject` in Listing 4.11 has a user agent identified with the PLASMA property `hasAgent`.

[3] `https://docs.inrupt.com/ess/latest/services/service-auditing/` (accessed on 21 December 2023)

Listing 4.10 Access control logs recorded in Beatriz's Pod.

```
<https://solidweb.me/besteves4/logs/AccessControl_Reject> a
 ↪  plasma:AccessControlLog ;
        dcterms:type as:Reject ;
        dcterms:issued "2023-11-12T15:34:04"^^xsd:dateTime ;
        as:summary "Access to data rejected" ;
        as:object <https://solidweb.me/besteves4/health/ehr> ;
        as:actor <https://solidweb.me/arya/profile/card#me> ;
        as:generator <https://example.com/App> ;
        dcterms:publisher
         ↪  <https://solidweb.me/besteves4/entities/LoggingProvider> .

<https://solidweb.me/besteves4/logs/AccessControl_Accept> a
 ↪  plasma:AccessControlLog ;
        dcterms:type as:Accept ;
        dcterms:issued "2023-11-12T15:43:09"^^xsd:dateTime ;
        as:summary "Access to data accepted" ;
        as:object <https://solidweb.me/besteves4/contacts/> ;
        as:actor <https://solidweb.me/arya/profile/card#me> ;
        as:generator <https://example.com/App> ;
        dcterms:publisher
         ↪  <https://solidweb.me/besteves4/entities/LoggingProvider> .

<https://example.com/App> a plasma:App .

<https://solidweb.me/besteves4/entities/LoggingProvider> a
 ↪  plasma:ServiceProvider ;
        plasma:serviceType plasma:LoggingService ;
        dpv:hasName "Entity G" ;
        dpv:hasAddress "Address of Entity G" ;
        dpv:hasContact "mailto:entity_g@mail.com" .
```

Listing 4.11 User registry of Beatriz's Pod.

```
<https://solidweb.me/besteves4/UserRegistry> a plasma:UserRegistry ;
    dcterms:description "Registry listing users" ;
    dcterms:created "2023-09-30T11:33:35"^^xsd:dateTime ;
    dcterms:modified "2023-10-07T12:39:50"^^xsd:dateTime ;
    dcterms:publisher
     ↪ <https://solidweb.me/besteves4/entities/UserMProvider> ;
    dpv:hasDataSubject
     ↪ <https://solidweb.me/besteves4/entities/DataSubject> ;
    plasma:hasUser <https://solidweb.me/besteves4/entities/DataSubject>,
        <https://solidweb.me/besteves4/entities/ReadUserA>,
        <https://solidweb.me/besteves4/entities/ReadUserB>,
        <https://solidweb.me/besteves4/entities/Admin> .

<https://solidweb.me/besteves4/entities/UserMProvider> a
 ↪ plasma:ServiceProvider ;
    dpv:hasName "Entity A" ;
    dpv:hasAddress "Address of Entity A" ;
    dpv:hasContact "mailto:entity_a@mail.com" .

<https://solidweb.me/besteves4/entities/DataSubject> a plasma:DataSubject,
 ↪ plasma:PodAdmin ;
    dpv:hasName "Data Subject" ;
    dpv:hasAddress "Address of Data Subject" ;
    dpv:hasContact "mailto:data_subject@mail.com" ;
    plasma:hasAgent <https://solidweb.me/besteves4/agents/AgentA> .

<https://solidweb.me/besteves4/agents/AgentA> a plasma:UserAgent .

<https://solidweb.me/besteves4/entities/ReadUserA> a plasma:ReadUser ;
    plasma:hasPolicy <https://solidweb.me/besteves4/requests/UserA> .

<https://solidweb.me/besteves4/entities/UserA> a plasma:DataRequest .

<https://solidweb.me/besteves4/entities/ReadUserB> a plasma:ReadUser ;
    plasma:hasPolicy <https://solidweb.me/besteves4/requests/UserB> ;
    plasma:hasPolicy <https://solidweb.me/besteves4/agreements/UserB> .

<https://solidweb.me/besteves4/agreements/UserB> a odrl:Agreement,
 ↪ plasma:UserConsent .

<https://solidweb.me/besteves4/entities/Admin> a plasma:AdminUser .
```

Finally, each of the requirements established in the previous paragraphs for the conformance of Pods, apps, services, users, and agents with PLASMA should be verified using a language for describing and validating RDF graphs. As such, SHACL can, not only, be used for the definition of data shapes recognised or supported by apps, services, or Pods, but also can act as a general tool to verify conformance with the conditions specified in this Section. SHACL was chosen for this conformance checking since it is a widely used and supported W3C Recommendation for *"validating RDF graphs against a set of conditions"* [Knublauch and Kontokostas, 2017]. As such, a SHACL validation tool takes a data graph and a shapes graph as input and produces a validation report that contains the outcomes of the validation. This validation report includes a set of all validation results, one for each shape being evaluated, which can be positive or negative, according to the result of the validation. SHACL shapes for all the previously mentioned conformance conditions were generated in this Thesis, and are available in PLASMA's online documentation and source code repository. Hence, conformance with the PLASMA specification relies on the proper modelling of policies, logs, registries and other metadata, using the terms suggested by the specification, i.e., for actual interoperability between Pods, apps, services, users, and agents. Moreover, SHACL validation reports can be used to assess which conformance conditions are not being met. While such reports are not enough to identify malicious actors, they can be used to advice users against using Pod, app, service, or agent providers which do not adhere to PLASMA's conformance conditions, and to ensure that the necessary metadata is available to the user in case such malicious uses are found, e.g., to be used by the data subject to lodge a complaint.

4.3.4 Vocabulary publication and maintenance

The vocabulary human-readable documentation and machine-readable file are available at https://w3id.org/plasma using content negotiation. The HTML documentation includes a description of the terms defined in PLASMA, which was conducted and validated with domain experts[4], diagrams with graphical representations of the several taxonomies included in the vocabulary, a detailed explanation of the conformance conditions that need to be adopted by Pods, apps, services, agents and users for them to be PLASMA-compliant, RDF examples of workflows that use PLASMA terms for specific scenarios, e.g., creating user policies or auditing Pods, and information related to legal compliance. The vocabulary documentation also includes metadata, such as the identity of the creators and publishers of the ontology, the dates of creation and last modification, or the version number.

The source code is hosted at https://w3id.org/plasma/repo, under the CC-BY-4.0 license. The repository can also be used by PLASMA implementers to suggest new inclusions to the vocabulary and to report bugs through GitHub Issues.

4.4 Exercising data subject rights with DPV

This Section describes the usage of vocabulary-based patterns to describe rights exercising metadata. Such patterns can be used by entities dealing with the handling of personal data to maintain

[4]This validation was performed with experts on the Solid technology, namely during the Short-Term Scientific Mission at the *KNoWS* group, as well as with the contacts made at *Inrupt*.

consistent records of data subject rights exercising activities, aligned with GDPR requirements. In a decentralised data system environment, these rights must also be fulfilled by data controllers, while notices and records of rights exercising activities can be kept by data subjects in their personal datastores for transparency and accountability.

4.4.1 Requirements to express rights-related activities

This Section outlines the motivation and identified requirements for the expression of information related to the exercising of data subject rights. This work was developed (and is already integrated) within the context of the DPVCG and was started with the main objectives of indicating (i) what rights exist (in particular within the framework of the GDPR), (ii) where such rights can be exercised, and (iii) what information needs to be recorded and maintained when a concrete instance of a right is being/was exercised.

As previously mentioned in Section 1.3.4 and represented in Figure 1.1, the focus of this Thesis is on the representation of information related to legislation on data protection in the European Union, in particular regarding the GDPR and related data subject rights, listed on Chapter III. Moreover, in Section 2.2.1, and in particular in Table 2.2, the privacy terms that need to be represented for such rights to be exercised by data subjects and fulfilled by data controllers.

Figure 4.9 illustrates the flows of information between a data subject and a data controller for the exercising of a right request, according to the GDPR. After sending a notice to the data subject confirming that the request was received, the controller must be able to identify the data subject in order to proceed with the request (Article 12.2, second sentence [2016b]). If the controller cannot identify the data subject, then the data subject must provide additional information to enable the controller to identify them (Article 11.2 [2016b]). If the controller disregards the request or has a justification for not fulfilling the right, then the data subject does not receive any information related to the right request (Article 12.2, second sentence [2016b]). In case the controller has a justification to delay the request due to its complexity or a high number of requests, then the controller has a 2-month extension to fulfil its duty (Article 12.3, second sentence [2016b]). Moreover, in case the request is unfounded or excessive, the controller can charge a fee and the data subject will get the information once this fee is paid (Article 12.5, first sentence [2016b]). As it is visible by the diagram, at any point if the data controller does not fulfil its duty then a GDPR breach occurs and the data subject does not receive their requested information.

From the analysis of these flows of information, a set of high-level concepts was proposed and adopted by the DPVCG (general concepts on Rights are modelled in the main DPV specification at `https://w3id.org/dpv#vocab-rights` and GDPR-specific ones in the DPV-GDPR extension at `https://w3id.org/dpv/legal/eu/gdpr#vocab-rights`[5]). Figure 4.10, adapted from Pandit et al. [2022], provides an overview of these concepts.

Thus, beyond modelling concepts for applicable `Rights` and `DataSubjectRights` (applicable only to data subjects), to indicate the association of concepts with a particular right, the `hasRight` property is also modelled in DPV. Additionally, to make a distinction between actionable and non-actionable rights, the `ActiveRight` and `PassiveRight` concepts were

[5]The concepts that have 'Beatriz Esteves' has a contributor are an outcome of this Thesis

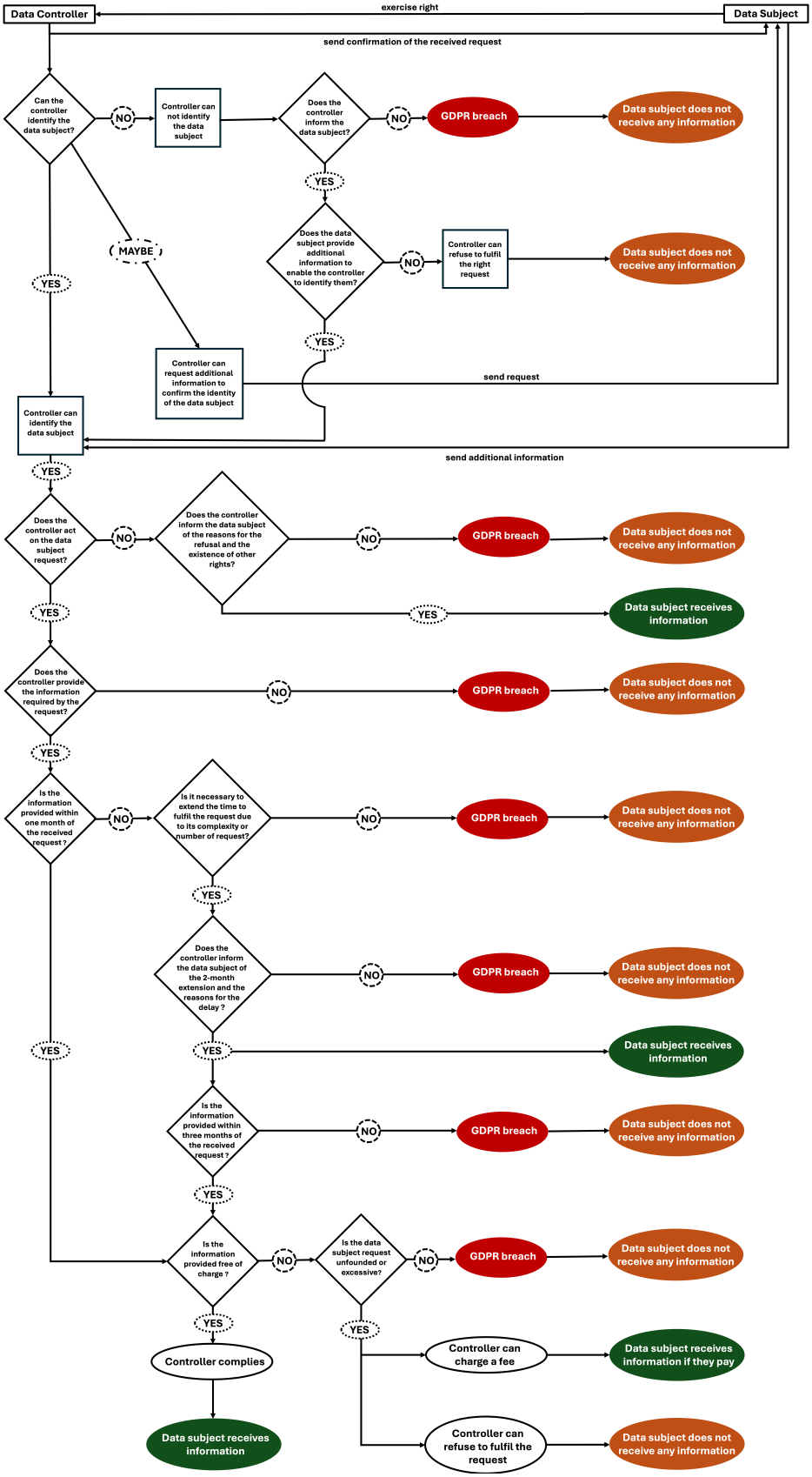


Figure 4.9: Flow diagram of GDPR data subject rights exercising, according to Article 12.

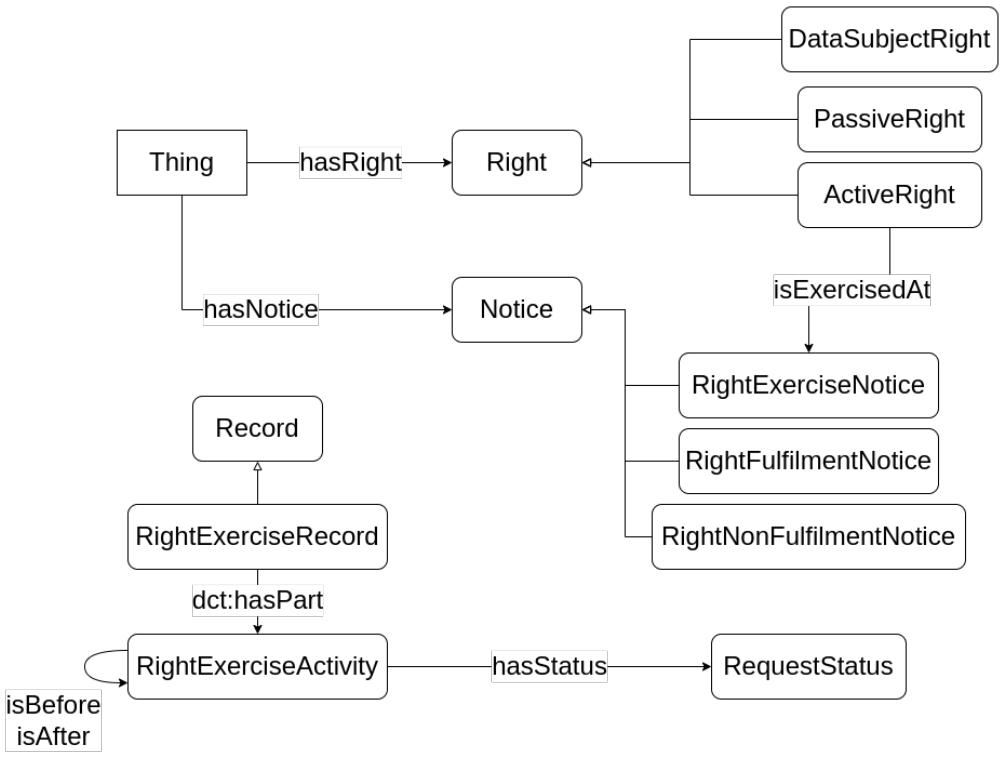


Figure 4.10: Core concepts of DPV's rights taxonomy, adapted from Pandit et al. [2022].

created to distinguish between rights that require an action to be taken for them to be exercised and rights that don't require any action and are always applicable.To fulfil the second objective of establishing where such active rights can be exercised, DPV's `isExercisedAt` property should be used to connect the right with the `RightExerciseNotice`. This notice provides contextual information regarding how to exercise a right. Specialised notice concepts for rights that can be fulfilled and those that cannot are modelled as `RightFulfilmentNotice` and `RightNonFulfilmentNotice`, respectively.

Moreover, to represent concrete records of rights being exercised, the `RightExerciseRecord` concept, specified as a subclass of DPV's `Record`, can be used to associate a particular request, or even distinct requests from the same data subject, with corresponding rights exercising activities, modelled as `RightExerciseActivity`, using the DCMI Metadata Terms `hasPart` property. Such activity instances should include metadata, e.g., timestamps, duration, or involved entities, to track the provenance of a particular right exercising process, from the request itself to its acknowledgement by the data controller and to the fulfilment or non-fulfilment of the right.

In order to justify a certain right exercise activity, a collection of justifications for the non-fulfilment, i.e., `RightNonFulfilmentJustification`, delay of fulfilment, i.e., `RightFulfilmentDelayJustification`, and exercise of rights, i.e., `RightExerciseJustification`, were modelled as subclasses of the `NonPerformanceJustification`, `DelayJustification`, and `ExerciseJustification` concepts, which were modelled to have generic justification concepts that can be used beyond the rights domain. `NotRequiredJustifications` are

also modelled for when a certain request is not required as it does not apply. Figure 4.11 contains the modelled justifications – they are modelled as generic justifications to be included in DPV, which are then extended in DPV-GDPR by referencing specific GDPR clauses. Moreover, the `dcterms:source` property will be used to connect the justification term with the GDPR provision that inspired its definition. These concepts have already been approved to be integrated into DPV and DPV-GDPR's outputs[6].

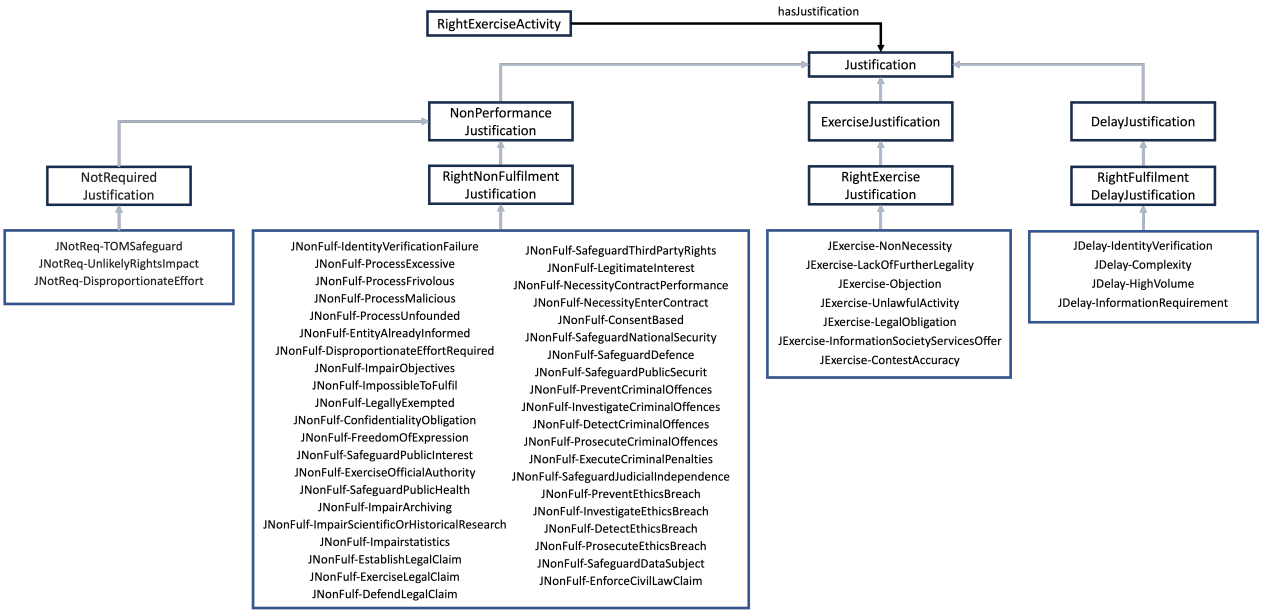


Figure 4.11: Justification concepts for the non-fulfilment, delay of fulfilment and exercise of rights.

Additionally, to track the status of rights exercising activities, a set of request statuses are modelled in DPV, including `RequestAccepted` for a request being accepted towards fulfilment, `RequestRejected` for a request being rejected towards non-fulfilment or `RequestRequiresAction` for a request requiring an action to be performed from another party, and the `isBefore` and `isAfter` concepts can be used to specify that a specific activity occurs before or after another activity.

While this modelling was inspired by the GDPR, the concepts are described in a jurisdiction-agnostic manner so that they can be used to tackle data protection regulations in different jurisdictions. For GDPR-specific rights, the `DataSubjectRight` concept is extended in DPV-GDPR with the data subject rights described in GDPR's Articles 13 to 22, as well as the rights to withdraw consent and to lodge a complaint with a supervisory authority, described in Articles 7.3 and 77. Moreover, notices for direct and indirect data collection, to fulfil the information requirements in Articles 13 and 14, for Subject Access Requests (SARs), described in Article 15, and for notifying recipients, necessary to fulfil the communication requirements of Articles 16, 17 and 18, are provided as GDPR-specific subclasses of `RightFulfilmentNotices`.

An overview of the aforementioned requirements and the intended purpose for modelling these concepts is presented in the ORSD illustrated in Table 4.9.

[6]Meeting notes of the DPVCG call where the concepts were accepted: https://w3id.org/dpv/meetings/meeting-2024-03-13.

Table 4.9: ORSD of the proposed model to express rights-related activities.

Vocabulary-based patterns for rights exercising activities	
1. Purpose	
The purpose of this model is the expression of rights-related activities, in particular focusing on data subject rights, such as the ones described in GDPR's Chapter III.	
2. Scope	
The scope of this model is limited to the expression of information related to the various steps of exercising data subject rights. In particular, the introduced concepts serve one of these purposes: (i) indicate what rights exist, (ii) express where such rights can be exercised, and (iii) record information related to concrete instances of rights that are being or were exercised.	
3. Implementation Language	
RDF, RDFS	
4. Intended End-Users	
Developers of Web services and applications, including decentralised storage solutions, that handle personal data.	
5. Intended Uses	
Use 1. Declaration of the existence of data subject rights when the usage and collection of personal data is performed by Web services providers and developers, including information on where they can be exercised. Use 2. Patterns for data subject rights that can be fulfilled. Use 3. Patterns for data subject rights that cannot be fulfilled, including justifications for non-fulfilment. Use 4. Fulfilment of data subject rights requests from specific data protection legislation, such as the GDPR.	
6. Ontology Requirements	
a. Non-Functional Requirements	
NFR 1. The concepts are either published online within DPVCG's outputs, following W3C's specification format, or are under discussion for being adopted by the same CG.	
b. Functional Requirements: Groups of Competency Questions	
CQRG1. Related to data subject rights	CQRG2. Related to GDPR
CQR1. What rights are applicable in a given context? CQR2. Where can the right be exercised? CQR3. How can the right be exercised? CQR4. What data is necessary to implement the right? CQR5. Which entity implements the right? CQR6. Which entity exercised the right? CQR7. When is the exercising activity occurring? CQR8. What is the status of the right request?	CQR9. Which data subject rights are applicable according to the legal basis used by the data controller? CQR10. Which provenance metadata must controllers provide when replying to data subject rights requests? CQR11. Which justification can be provided to not fulfil, delay or exercise a request?

4.4.2 Vocabulary-based patterns for rights exercising activities

This Section outlines the usage of DPV for the expression of rights exercising activities, specifically related to the data subject rights described in GDPR's Chapter III. Accordingly, the *"controller shall take appropriate measures to provide any information referred to in Articles 13 and 14 and any communication under Articles 15 to 22 [...] to the data subject in a concise, transparent, intelligible and easily accessible form"* and if *"the data subject makes the request by electronic form means, the information shall be provided by electronic means where possible, unless otherwise requested by the data subject"* [2016b]. Consequently, in this Thesis, personal data processing-related vocabularies such as DPV, and other semantic metadata vocabularies such as DCMI Metadata Terms and DCAT, are used for the establishment of structured notices and records of activities, which promote the fulfilment of the transparency and machine-readability requirements previously described. Additionally, by storing such information in decentralised personal datastores, the accessibility requirement can also be satisfied.

Listing 4.12 provides an example of a personal data handling activity, which can be used to express information regarding the what, how, where, who, and why personal data is being processed, as well as what rights exist. The example provides information regarding the type of personal data, `pd:EmailAddress`, being processed by the `ex:DataController`, the purpose and legal basis used for the processing to occur, `dpv:ServiceProvision` and `eu-gdpr:A6-1-a`, and the applicable GDPR data subject rights. This pattern can be followed by data controllers to express which rights are applicable, including rights beyond the ones in the GDPR, e.g., EU's fundamental rights and the rights depicted in other EU regulations or in other jurisdictions.

Listing 4.12 Personal data handling activity example which includes information regarding the applicable rights.

```
ex:ProcessEmailForServiceProvision a dpv:PersonalDataHandling ;
    dpv:hasDataController ex:DataController ;
    dpv:hasPersonalData pd:EmailAddress ;
    dpv:hasProcessing dpv:Collect, dpv:Use ;
    dpv:hasPurpose dpv:ServiceProvision ;
    dpv:hasLegalBasis eu-gdpr:A6-1-a ;
    dpv:hasRight eu-gdpr:A7-3, eu-gdpr:A13, eu-gdpr:A14, eu-gdpr:A15,
    ↪ eu-gdpr:A16, eu-gdpr:A17, eu-gdpr:A18, eu-gdpr:A20, eu-gdpr:A22,
    ↪ eu-gdpr:A77 .
ex:DataController a dpv:DataController .
```

Moreover, such declarations of applicable rights should include a notice of where they can be exercised. Thus, as described in the previous Section, DPV's `isExercisedAt` property should be used to connect rights with information on where to exercise it. Such information should be provided through a `RightExerciseNotice`, along with other rights exercising metadata.

Listing 4.13 provides a notice with information on where to exercise the GDPR's right of access related to personal data being processed by the `ex:DataController`. This notice uses the `dpv:hasRight` property to indicate which rights can be exercised, beyond the already mentioned access right, and the `foaf:page` property to express the precise Web page where the right can be exercised. Additionally, DPV's `hasDataController` and

`isImplementedByEntity` properties are used to define who is the controller responsible for the personal data being processed and who is the entity implementing the service/platform where the rights are exercised – in most cases this entity will probably coincide with the data controller, however, for transparency and accountability purposes, it is reasonable to express both terms. Other entity-related properties can also be used to personalise the notice, e.g., if a notice is personalised for a specific data subject, then `dpv:hasDataSubject` can be used to connect the notice with the individual data subject. Furthermore, personal data handling instances can also be used to express 'data processing bundles' that need to be provided in order to fulfil the data subjects' rights. As previously mentioned, this can be done using DPV's taxonomies of purposes, personal data categories, processing operations, etc, e.g., in Listing 4.13, an account identifier is required by the data controller for identity verification purposes in order to fulfil the data subject's rights described in GDPR's Articles 7.3, 15, 16, 17 and 20. Other information might be needed to be communicated to the user, for instance, information on payments as Article 12.5(a) states that a fee *"taking into account the administrative costs of providing the information or communication or taking the action requested"* might be requested by the data controller in case the data subject's request is unfounded, excessive or repetitive. Such information can also be expressed in personal data handling instances including ODRL policies with duties constrained with an `odrl:payAmount` left operand.

Listing 4.13 GDPR Article 15's right of access exercise notice, including information on where to exercise the right and on necessary data to fulfil the right.

```
ex:RightToAccess a eu-gdpr:A15 ;
    dpv:isExercisedAt ex:RightExercisePoint .

ex:RightExercisePoint a dpv:RightExerciseNotice ;
    dpv:hasRight eu-gdpr:A7-3, eu-gdpr:A15, eu-gdpr:A16, eu-gdpr:A17,
     ↪ eu-gdpr:A20 ;
    dpv:hasDataController ex:DataController ;
    dpv:isImplementedByEntity ex:DataController ;
    foaf:page <https://example.com/DataController/RightExercisePoint> ;
    dpv:hasPersonalDataHandling [
        a dpv:PersonalDataHandling ;
        dpv:hasPurpose dpv:IdentityVerification ;
        dpv:hasPersonalData pd:AccountIdentifier ;
        dpv:hasProcessing dpv:Collect, dpv:Store ;
        dpv:hasRecipientDataController ex:DataController ] .
```

In addition to notices expressing what rights are applicable and where they can be exercised, provenance metadata must be kept when actual instances of the rights are exercised, throughout the whole process, from initiating a request to rejecting or fulfilling it. As such, DPV's `RightExerciseActivity` can be used in conjunction with the W3C's PROV-O recommendation [Lebo et al., 2013] to track the provenance of a right exercising activity instance. Using this standard, provenance information, regarding the entities whom the request is associated with, i.e., `prov:wasAssociatedWith`, or what data/notice was generated, i.e., `prov:generated`, by the right exercise activity, can be represented. `prov:actedOnBehalfOf` can also be used to represent delegation or representation, for instance when a parent exercises a right on behalf of its child. Temporal information, descriptions and identifiers of the activities and their

creators/publishers can be recorded using DCMI Metadata Terms. Connections to the previous or following activities can be done using the `dpv:isAfter` and `dpv:isBefore` concepts, respectively. Moreover, to track the status of a right exercising activity, DPV's taxonomy of request statuses concepts can be used.

Figure 4.12 illustrates the sequence in which these concepts occur. Once the request is initiated, it should be then acknowledged by the entity implementing it and either accepted towards fulfilment or rejected towards non-fulfilment. Additionally, after being rejected, the entity fulfilling the request can also require further action from the requester (e.g., request additional data to be able to fulfil the request), which can delay the acceptance or rejection of the request, and after the required action is performed, the request can either be accepted towards fulfilment or get rejected again towards non-fulfilment or towards asking again for further action.

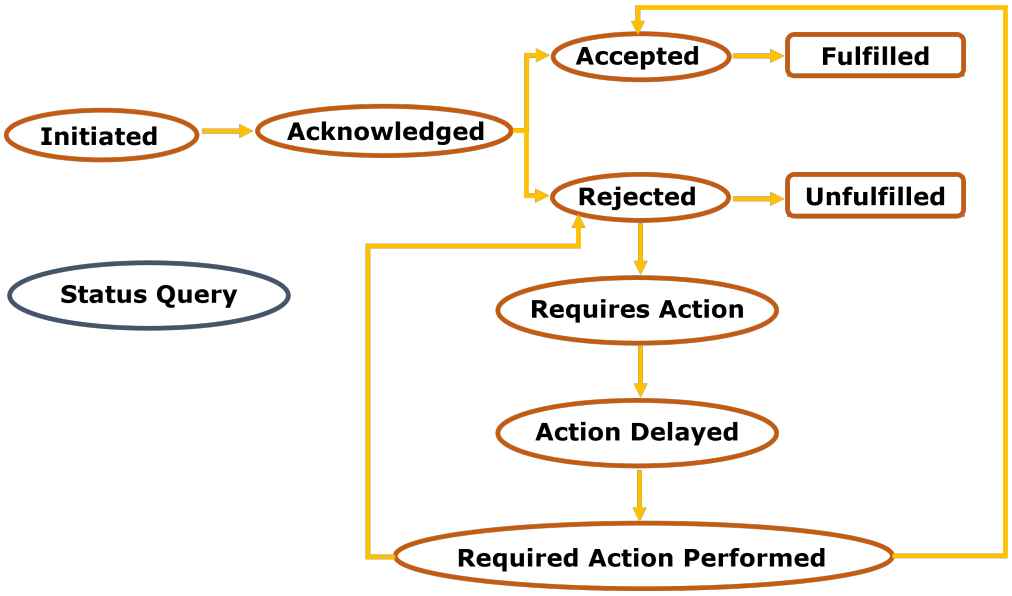


Figure 4.12: DPV's concepts to model the status of a request.

Listing 4.14 illustrates a record of the right exercising activities related to a GDPR right of access request and acknowledgement of said request. The activity associated with the start of the request has the status `dpv:RequestInitiated`, the data subject is identified using the `dpv:hasDataSubject` and the recipient of the request, a data controller, using the `dpv:hasRecipientDataController`. Furthermore, a personal data handling instance can be used to express what personal data needs to be processed for the fulfilment of the right and DPV's `hasScope` to specify the scope of the request, e.g., the data subject only wants to access data processed for service provision purposes or only data processed during 2022. Following the start of the request, the controller acknowledges the request, a right exercising activity which has `dpv:RequestAcknowledged` status and the recipient is the data subject that initiated the request.

Listing 4.15 illustrates the follow-up to Listing 4.14 – the request was rejected due to the data controller not being able to identify the data subject, `JNonFulf-IdentityVerificationFailure`. Table 4.10 contains the list of modelled

Listing 4.14 Record of GDPR right of access request and acknowledgement activities.

```
ex:DataSubject a dpv:DataSubject .
ex:DataSubjectUsername a pd:AccountIdentifier ;
    dpv:hasDataSubject ex:DataSubject .

ex:SARequest a dpv:RightExerciseActivity, prov:Activity ;
    dcterms:description "Data Subject makes a GDPR right of access
     ↪ request" ;
    dpv:hasRight eu-gdpr:A15 ;
    dpv:isExercisedAt ex:RightExercisePoint ;
    prov:wasAssociatedWith ex:DataSubject ;
    dpv:hasDataSubject ex:DataSubject ;
    dpv:hasRecipientDataController ex:DataController ;
    dcterms:date "2023-11-02T11:08:05"^^xsd:dateTime ;
    dpv:hasStatus dpv:RequestInitiated ;
    dpv:hasPersonalDataHandling [
        a dpv:PersonalDataHandling ;
        dpv:hasPurpose dpv:IdentityVerification ;
        dpv:hasPersonalData ex:DataSubjectUsername ;
        dpv:hasProcessing dpv:Collect, dpv:Store ] ;
    dpv:hasScope [
        dpv:hasPurpose dpv:ServiceProvision ;
        dpv:hasDuration [
            time:hasBeginning "2022-01-01T00:00:00"^^xsd:dateTime ;
            time:hasEnd "2022-12-31T23:59:59"^^xsd:dateTime ] ] .

ex:SARAcknowledged a dpv:RightExerciseActivity, prov:Activity ;
    dcterms:description "Data controller acknowledges the request" ;
    dcterms:date "2023-11-02T15:55:10"^^xsd:dateTime ;
    prov:wasAssociatedWith ex:DataController ;
    dpv:hasRecipient ex:DataSubjectUsername ;
    dpv:hasStatus dpv:RequestAcknowledged ;
    dpv:isAfter ex:SARequest .
```

right non-fulfilment justifications and their labels, which can be used by controllers for the purpose of justifying such requests. As such the data controller requires further information from the data subject to be able to proceed with the request. Such right exercise activity is identified with the `dpv:RequestRequiresAction` status and contains a personal data handling activity instance expressing the information that the data subject needs to provide for the right exercise to continue. Afterwards, a right exercise activity associated with the data subject and with a `dpv:RequestRequiredActionPerformed` status is recorded with the information that the data subject provided.

Listing 4.15 Record of data controller requesting further information to fulfil the data subject's SAR and of the data subject providing the controller with said information.

```
ex:SARRejected a dpv:RightExerciseActivity, prov:Activity;
    dcterms:description "Data controller rejects the request" ;
    dcterms:date "2023-11-02T15:57:31"^^xsd:dateTime ;
    prov:wasAssociatedWith ex:DataController ;
    dpv:hasRecipient ex:DataSubjectUsername ;
    dpv:hasStatus dpv:RequestRejected ;
    dpv:hasJustification justif:JNonFulf-IdentityVerificationFailure ;
    dpv:isAfter ex:SARAcknowledged .

ex:SARRequiresAction a dpv:RightExerciseActivity, prov:Activity ;
    dcterms:description "Data controller requires further actions" ;
    dcterms:date "2023-11-02T16:09:21"^^xsd:dateTime ;
    prov:wasAssociatedWith ex:DataController ;
    dpv:hasRecipient ex:DataSubjectUsername ;
    dpv:hasStatus dpv:RequestRequiresAction ;
    dpv:hasJustification justif:JNonFulf-IdentityVerificationFailure ;
    dpv:hasPersonalDataHandling [
        dpv:hasPersonalData pd:OfficialID ;
        dpv:hasProcessing dpv:MakeAvailable ;
        dpv:hasPurpose dpv:IdentityVerification ;
        dpv:hasRecipientDataController ex:DataController ;
        dpv:isImplementedByEntity ex:DataSubjectUsername ] ;
    dpv:isAfter ex:SARRejected .

ex:DataSubjectOfficialID a pd:OfficialID ;
    dpv:hasDataSubject ex:DataSubject .

ex:SARActionPerformed a dpv:RightExerciseActivity, prov:Activity ;
    dcterms:description "Data Subject provides required information" ;
    dcterms:date "2023-11-02T17:20:42"^^xsd:dateTime ;
    prov:wasAssociatedWith ex:DataSubject ;
    dpv:hasStatus dpv:RequestRequiredActionPerformed ;
    dpv:hasPersonalDataHandling [
        dpv:hasPersonalData ex:DataSubjectOfficialID ;
        dpv:hasProcessing dpv:MakeAvailable ;
        dpv:hasPurpose dpv:IdentityVerification ;
        dpv:hasRecipientDataController ex:DataController ;
        dpv:isImplementedByEntity ex:DataSubjectUsername ] ;
    dpv:isAfter ex:SARRequiresAction .
```

Table 4.10: Justifications for non-fulfilment of GDPR's data subject rights.

Term	Label	GDPR Article(s)
JNonFulf-IdentityVerificationFailure	Justification that the process could not be fulfilled or was not successful because identity verification failed	12.2
JNonFulf-ProcessExcessive	Request was excessive in scope	12.5
JNonFulf-ProcessFrivolous	Request was frivolous in scope	12.5
JNonFulf-ProcessMalicious	Request was malicious in scope	12.5
JNonFulf-ProcessUnfounded	Request was unfounded in scope	12.5
JNonFulf-EntityAlreadyInformed	Data subject already has been provided with this information	13.4, 14.5(a)
JNonFulf-ImpairObjectives	Fulfilment would cause impairment to processing	14.5(b)
JNonFulf-DisproportionateEffortRequired	Fulfilment would require extraordinary effort	14.5(b), 19
JNonFulf-ImpossibleToFulfil	Fulfilment would be impossible	14.5(b), 19
JNonFulf-LegallyExempted	Fulfilment not needed as it falls under legal exemption	14.5(c), 17.3(b), 22.2(b)
JNonFulf-ConfidentialityObligation	Fulfilment would compromise existing confidentiality obligations	14.5(d)
JNonFulf-FreedomOfExpression	Fulfilment would interfere with the right of freedom of expression and information of others	17.3(a)
JNonFulf-SafeguardPublicInterest	Fulfilment would interfere with necessary tasks carried out for public interest	17.3(b), 21.6, 23.1(e)
JNonFulf-ExerciseOfficialAuthority	Fulfilment would interfere with the exercise of official authorities	17.3(b), 20.3, 23.1(h)
JNonFulf-SafeguardPublicHealth	Fulfilment would interfere with necessary tasks carried out for public health reasons	17.3(c)
JNonFulf-ImpairArchiving	Fulfilment would compromise or hinder archiving purposes	17.3(d)
JNonFulf-ImpairScientificOrHistoricalResearch	Fulfilment would impair scientific or historical research	17.3(d)
JNonFulf-ImpairStatistics	Fulfilment would interfere with official statistics	17.3(d)
JNonFulf-EstablishLegalClaim	Fulfilment would interfere with the establishment of legal claims	17.3(e), 21.1
JNonFulf-ExerciseLegalClaim	Fulfilment would interfere with the exercise of legal claims	17.3(e), 21.1
JNonFulf-DefendLegalClaim	Fulfilment would interfere with the defence of legal claims	17.3(e), 21.1
JNonFulf-SafeguardThirdPartyRights	Fulfilment would adversely affect the rights of other data subjects or third parties	20.4, 23.1(i)
JNonFulf-LegitimateInterest	Fulfilment would interfere with the legitimate interest of the controller which overrides the interests or rights of the data subject	21.1
JNonFulf-NecessityContractPerformance	Fulfilment would interfere with the performance of a contract	22.2(a)
JNonFulf-NecessityEnterContract	Fulfilment would interfere with the necessity of entering into a contract	22.2(a)
JNonFulf-ConsentBased	Fulfilment not necessary as processing is based on explicit consent	22.2(c)
JNonFulf-SafeguardNationalSecurity	Fulfilment would pose a threat to safeguard national security	23.1(a)
JNonFulf-SafeguardDefence	Fulfilment would pose a threat to safeguard defence	23.1(b)
JNonFulf-SafeguardPublicSecurity	Fulfilment would pose a threat to safeguard public security	23.1(c)
JNonFulf-PreventCriminalOffences	Fulfilment would interfere with the prevention of criminal offences	23.1(d)
JNonFulf-InvestigateCriminalOffences	Fulfilment would interfere with the investigation of criminal offences	23.1(d)
JNonFulf-DetectCriminalOffences	Fulfilment would interfere with the detection of criminal offences	23.1(d)
JNonFulf-ProsecuteCriminalOffences	Fulfilment would interfere with the prosecution of criminal offences	23.1(d)
JNonFulf-ExecuteCriminalPenalties	Fulfilment would interfere with the execution of criminal penalties	23.1(d)
JNonFulf-SafeguardJudicialIndependence	Fulfilment would pose a threat to safeguard judicial independence or proceedings	23.1(f)
JNonFulf-PreventEthicsBreach	Fulfilment would interfere with the prevention of breaches of ethics for regulated professions	23.1(g)
JNonFulf-InvestigateEthicsBreach	Fulfilment would interfere with the investigation of breaches of ethics for regulated professions	23.1(g)
JNonFulf-DetectEthicsBreach	Fulfilment would interfere with the detection of breaches of ethics for regulated professions	23.1(g)
JNonFulf-ProsecuteEthicsBreach	Fulfilment would interfere with the prosecution of breaches of ethics for regulated professions	23.1(g)
JNonFulf-SafeguardDataSubject	Fulfilment would interfere with the protection of the data subject	23.1(i)
JNonFulf-EnforceCivilLawClaim	Fulfilment would interfere with the enforcement of civil law claims	23.1(j)

Listing 4.16 illustrates the follow-up to Listing 4.15 – the request was accepted and fulfilled by the data controller and as such the data controller provides the data subject with a notice of the fulfilment of GDPR's Art.15, modelled as a `eu-gdpr:SARNotice`, and a copy of the data whose access was requested, modelled as a `dcat:Dataset`. Beyond temporal information and providing the location of the notice, a personal data handling instance can be used to provide additional information to the data subject – in the case of this SAR notice, the data subject is also notified of the type of personal data being processed, the purpose for the processing, the time period when the data was processed and the additional data subject rights that can be exercised.

Listing 4.16 Record of the acceptance and fulfilment of the request and respective `SARNotice`.

```
ex:SARAccepted a dpv:RightExerciseActivity, prov:Activity ;
    dcterms:description "Request accepted by data controller towards
     ↪ fulfilment" ;
    dcterms:date "2023-11-03T08:15:04"^^xsd:dateTime ;
    prov:wasAssociatedWith ex:DataController ;
    dpv:hasRecipient ex:DataSubjectUsername ;
    dpv:hasStatus dpv:RequestAccepted ;
    dpv:isAfter ex:SARActionPerformed .

ex:SARFulfilled a dpv:RightExerciseActivity, prov:Activity ;
    dcterms:description "Request fulfilled by data controller" ;
    dcterms:date "2023-11-03T08:37:25"^^xsd:dateTime ;
    prov:wasAssociatedWith ex:DataController ;
    dpv:hasRecipient ex:DataSubjectUsername ;
    dpv:hasStatus dpv:RequestFulfilled ;
    prov:generated ex:DataCopy, ex:SARNotice_Username ;
    dpv:isAfter ex:SARAccepted .

ex:SARNotice_Username a eu-gdpr:SARNotice ;
    dcterms:date "2023-11-03T08:31:51"^^xsd:dateTime ;
    foaf:page <https://example.com/DataController/SARNotice_Username> ;
    dpv:hasPersonalDataHandling [
        dpv:hasPersonalData pd:EmailAddress ;
        dpv:hasPurpose dpv:ServiceProvision ;
        dpv:hasDuration [
            time:hasBeginning "2022-01-01T00:00:00"^^xsd:dateTime ;
            time:hasEnd "2022-12-31T23:59:59"^^xsd:dateTime ] ;
        dpv:hasRight eu-gdpr:A16, eu-gdpr:A17, eu-gdpr:A18, eu-gdpr:A21,
         ↪ eu-gdpr:A77 ] .

ex:DataCopy a dcat:Dataset ;
    dcterms:format
     ↪ <https://www.iana.org/assignments/media-types/text/csv> ;
    dcterms:accessRights access-right:c_16ec ; # restricted access
    dcterms:issued "2023-11-03T08:35:42"^^xsd:dateTime ;
    dcterms:valid "2023-12-03T08:35:42"^^xsd:dateTime ;
    dcat:landingPage <https://example.com/Username/SAR_DataCopy> .
```

Moreover, DCAT [Albertoni et al., 2020] can be used to model resources beyond notices, for instance, a copy of the personal data, in the case of a right of access request according to the

Table 4.11: Justifications for the exercise of GDPR's data subject rights.

Term	Label	GDPR Article(s)
JExercise-NonNecessity	Processing no longer necessary for the specified purposes	17.1(a), 18.1(c)
JExercise-LackOfFurtherLegality	Processing no longer necessary due to a lack of further legality of the legal basis of specified context	17.1(b)
JExercise-Objection	Data subject objected to the processing	17.1(c), 18.1(d)
JExercise-UnlawfulActivity	Personal data unlawfully processed	17.1(d), 18.1(b)
JExercise-LegalObligation	Compliance with a legal obligation	17.1(e)
JExercise-InformationSocietyServicesOffer	Personal data collected in relation to the offer of information society services	17.1(f)
JExercise-ContestAccuracy	Accuracy of personal data contested by the data subject	18.1(a)

GDPR, as in Listing 4.16. Information regarding the format, validity and dataset provision/-download location can be attached to the dataset representation using the `dcterms:format`, `dcterms:valid` and `dcat:landingPage` properties, respectively. Additionally, DCAT – and also the Data on the Web Best Practices Recommendation [Lóscio et al., 2017] – promotes the usage of ODRL to express license and rights statements, by linking the dataset with an ODRL policy using the `odrl:hasPolicy` property, or the usage of DCMI Metadata Terms' `license`, `accessRights` or `rights` properties to link datasets with licenses, access rights statements or other types of rights statements, e.g., copyrights, respectively. In the case of using the latter, controlled vocabularies such as COAR's Access Rights vocabulary [Apollaro et al., 2022], used in Listing 4.16 to restrict access only to the data subject, or the Named Authority List of Access rights from the Publications Office of the European Union [2023] can be used to express 'high-level' access control statements, e.g., embargoed, restricted or open access. As such, by using DCAT and a policy expression vocabulary, i.e., ODRL, the user can easily control the transition of a compliance log and respective metadata, as proposed in PLASMA, from being only accessible by the user to be accessible by external auditors.

In this Section, the GDPR's Right of Access is used as an example to showcase how to model notices and right exercising activities using DPV, DCMI Metadata Terms, PROV-O and DCAT. However, a similar pattern can be followed by data controllers to fulfil the other rights as in most cases the only substantial change would be the notice concept that needs to be used for a particular right instance, e.g., `eu-gdpr:DirectDataCollectionNotice` for the right fulfilment notice related to GDPR's Article 13, `eu-gdpr:IndirectDataCollectionNotice` for the right fulfilment notice related to GDPR's Article 14 or `eu-gdpr:RightsRecipientsNotice` for the right fulfilment notice related to GDPR's Article 19. Moreover, some GDPR rights, such as the right to erasure in Article 17 and the right to restriction of processing in Article 18, also require the data subject to provide a justification for their request to be fulfilled by the data controller. As such, a collection of justifications for the exercise of data subject rights, extracted from the GDPR and illustrated in Table 4.11, was also modelled (as subclasses of a high-level `RightExerciseJustification` concept).

4.4.3 Justifications publication and maintenance

As discussed in the previous sections, most rights-related concepts were already proposed and are integrated into DPV and DPV-GDPR. The justifications concepts have already been approved to be accepted, but are still to be published in DPVCG's outputs. As such, the justifications vocabulary human-readable documentation and machine-readable file are available at https://w3id.org/people/besteves/justifications using content negotiation. Once the justification terms are fully integrated into DPV, this documentation will be archived as the maintenance of the terms will be performed as part of the DPVCG's activities.

The HTML documentation includes a description of the defined terms, which was conducted and validated with domain experts[7], diagrams with graphical representations of the several taxonomies used in the vocabulary patterns specific in this section, and RDF examples that use the defined terms to express rights-related activities. The vocabulary documentation also includes metadata, such as the identity of the creators and publishers of the ontology, the dates of creation and last modification or the version number. The source code is hosted at https://w3id.org/people/besteves/justifications/repo, under the CC-BY-4.0 license. The repository can also be used by implementers to suggest new inclusions to the vocabulary and to report bugs through GitHub Issues.

Additionally, an auxiliary webpage, openly available at https://w3id.org/people/besteves/rights, provides guidelines and further examples on how to use DPV and the developed `Justification` terms to model the vocabulary-based patterns for rights-related activities defined in this Section. The source code is hosted at https://w3id.org/people/besteves/rights/repo, under the CC-BY-4.0 license.

4.5 Ontology evaluation

This Section describes the results of the ontologies evaluation, including the detection of common pitfalls with OOPS![8] [Poveda-Villalón et al., 2014], alignment with FAIR principles with FOOPS![9] [Garijo et al., 2021] and validation of competency questions with SPARQL queries [Harris and Seaborne, 2013].

In terms of quality evaluation, the OOPS! tool was used to evaluate PLASMA and OAC in order to detect common errors in ontology development, such as missing domain or range properties or missing human-readable annotations. Both evaluations did not detect any critical (issues affecting the ontology consistency, reasoning, or applicability) or important (issues not critical for ontology functionality but that should be corrected) pitfalls. Furthermore, FOOPS! was used to evaluate the alignment of the developed vocabularies with the FAIR (Findable, Accessible, Interoperable and Reusable) principles. Additionally, the vocabularies that were reused in this Section, e.g., DPV, ODRL, or ActivityStreams, were also evaluated with FOOPS! for comparison. Table 4.12 presents the results of this evaluation.

[7]This validation was performed with legal and ontology engineering experts in the context of W3C's DPV community group.

[8]The OOPS! tool is available at https://oops.linkeddata.es/ (accessed on 14 June 2023).

[9]The FOOPS! tool is available at https://w3id.org/foops (accessed on 30 November 2023).

Table 4.12: Evaluation of the alignment of the developed and reused vocabularies with FAIR principles using FOOPS!.

Ontology	FOOPS! score	Findable	Accessible	Interoperable	Reusable
OAC	91%	8/9	2/3	3/3	8.83/9
PLASMA	91%	8/9	2/3	3/3	8.83/9
DPV	64%	5.33/9	2/3	3/3	4.92/9
ODRL	64%	4.5/9	3/3	3/3	4.75/9
DCAT	64%	4.33/9	3/3	3/3	5.14/9
ACP	52%	5.33/9	2/3	2/3	3.12/9
ActivityStreams	19%	2/9	1.5/3	0/3	1/9
ACL	2%	0/9	0.5/3	0/3	0/9
DCMI	2%	0/9	0.5/3	0/3	0/9

Both PLASMA and OAC obtained a good score in all FAIR aspects. In terms of improvements, both vocabularies will be submitted to LOV (Linked Open Vocabularies), a public registry of ontologies that includes ontology metadata related to its terms and the creators/developers of the ontologies [Vandenbussche et al., 2017]. This will improve both the findability and the accessibility of the vocabularies. Using Table 4.12, it can be observed that both PLASMA and OAC rate much higher than other reused ontologies in terms of findability and reusability as the used URIs and version URIs are persistent and resolvable, both include the recommended ontology and ontology terms metadata, a resolvable data usage license and provenance metadata.

Additionally, SPARQL queries are used to assess whether the developed models satisfy the competency questions, presented in Tables 4.4, 4.7 and 4.9, which were used to guide the development of the vocabularies in order to fulfil the identified requirements.

Table 4.13 presents the SPARQL queries drafted to fulfil OAC's competency questions presented in Table 4.4. The presented work demonstrates that OAC satisfies the identified requirements of answering competency questions regarding the modelling of policies representing personal preferences, requests of data for particular purposes, and agreements of data access, including contextual information. In particular, these competency questions showcase OAC's capabilities for allowing the definition of different user preferences as policies, the specification of permissions and prohibitions at an arbitrary level of granularity, the identification and resolution of conflicting policies, and the usage of legally-aligned concepts, while also supporting users with easy access to the policies used for granting/denying access to data for internal and/or external introspection, thus fulfilling OAC's requirements outlined in Section 4.2.1.

Moreover, Table 4.14 lists concepts that can be used to answer the competency questions identified in PLASMA's ORSD, available in Table 4.7. Listing 4.17 illustrates how these concepts can be used in SPARQL queries to fulfil the *"CQP1. Which Pod management data is stored in the Pod?"* and *"CQP4. What policy describes the data access requirements of a certain app or service?"* competency questions. A similar exercise can be done for the remaining competency questions. The presented work demonstrates that PLASMA satisfies the requirements identified in Section 4.3.1 of answering competency questions regarding the representation of information related to entities, infrastructure,

Table 4.13: Validation of OAC's competency questions with SPARQL queries.

CQO*	SPARQL query
CQO1	SELECT ?policy WHERE { ?policy_uri a ?policy . ?policy_uri odrl:uid ?policy_id . }
CQO2	SELECT ?action WHERE { ?policy_uri a ?policy . ?policy_uri odrl:uid ?policy_id . ?policy_uri odrl:permission|odrl:prohibition ?rule . ?rule odrl:action ?action . }
CQO3	SELECT ?data WHERE { ?policy_uri a ?policy . ?policy_uri odrl:uid ?policy_id . ?policy_uri odrl:permission|odrl:prohibition ?rule . ?rule odrl:target ?data . }
CQO4	SELECT ?constraint WHERE { ?policy_uri a ?policy . ?policy_uri odrl:uid ?policy_id . ?policy_uri odrl:permission ?rule . ?rule odrl:constraint ?constraint . }
CQO5	SELECT ?assigner ?assignee WHERE { ?policy_uri a ?policy . ?policy_uri odrl:uid ?policy_id . ?policy_uri odrl:permission|odrl:prohibition ?rule . OPTIONAL { ?rule odrl:assigner ?assigner } . OPTIONAL { ?rule odrl:assignee ?assignee } . }
CQO6	SELECT ?conflict_term WHERE { ?policy_uri a ?policy . ?policy_uri odrl:uid ?policy_id . ?policy_uri odrl:conflict ?conflict_term . }
CQO7	SELECT ?context WHERE { ?policy_uri a ?policy . ?policy_uri odrl:uid ?policy_id . ?policy_uri odrl:permission|odrl:prohibition ?rule . ?rule dpv:hasContext ?context . }
CQO8	SELECT ?legal_basis WHERE { ?policy_uri a ?policy . ?policy_uri odrl:uid ?policy_id . ?policy_uri odrl:permission|odrl:prohibition ?rule . ?rule odrl:constraint ?constraint . ?constraint odrl:leftOperand oac:LegalBasis . ?constraint odrl:rightOperand ?legal_basis . }
CQO9	SELECT ?entity ?address ?contact ?name WHERE { ?entity a oac:Entity . ?entity dpv:hasAddress ?address . ?entity dpv:hasContact ?contact . ?entity dpv:hasName ?name . }

and processes in Solid, the modelling of information related to legal roles in a jurisdiction-agnostic manner, and the definition of patterns to express apps and services policies, data usage logs and registries of data, schemas, apps, services, entities, and authorisations for convenient access to information. The developed SHACL shapes also satisfy the identified requirements by ensuring compliance with PLASMA's conformance conditions described in detail in Section 4.3.3.

Table 4.14: Concepts in PLASMA and other vocabularies for answering competency questions defined in Table 4.7.

CQP*	Concepts
CQP1	`PodManagementData`, `hasProvider`, `hasDeveloper`, `implementedSolidPlatform`, `implementedSolidSpecification`
CQP2	`DataLog`, `DataProvider`, `ResourceMetadata`, `DataAgreement`
CQP3	`Policy`, `Log`, `UserData`, `AppData`, `ServiceData`, `PodManagementData`, `SchemaData`
CQP4	`DataRequest`, `AppManifest`, `ServiceManifest`
CQP5	`InfrastructureProvider`, `PodProvider`, `PodDeveloper`, `SolidPlatformProvider`, `SolidPlatformProvider`
CQP6	`DataLog`, `DataProvider`, `dcat:landingPage`, `dcat:distribution`, `dcat:accessURL`, `dcat:mediaType`
CQP7	`AccessControlRegistry`, `DataRegistry`, `DataSchemaRegistry`, `PolicyRegistry`, `AppRegistry`, `ServiceRegistry`, `UserRegistry`
CQP8	`dpv:hasName`, `dpv:hasAddress`, `dpv:hasContact`
CQP9	`InfrastructureProvider`, `PodProvider`, `SolidPlatformProvider`, `dpv:Purpose`, `dpv:LegalBasis`, `Log`, `Notice`

Listing 4.17 SPARQL queries to validate PLASMA's CQP1 and CQP4.

```
SELECT ?provider ?developer ?platform ?spec WHERE {
    ?pod_data a plasma:PodManagementData .
    ?pod_data plasma:hasProvider ?provider .
    ?pod_data plasma:hasDeveloper ?developer .
    ?pod_data plasma:implementedSolidPlatform ?platform .
    ?pod_data plasma:implementedSolidSpecification ?spec . }

SELECT ?app ?appmanifest ?policy WHERE {
    ?app a plasma:App .
    ?app plasma:hasAppManifest ?appmanifest .
    ?appmanifest a plasma:AppManifest .
    ?appmanifest odrl:hasPolicy ?policy . }
```

Finally, Table 4.15 presents the SPARQL queries drafted to fulfil the competency questions of the proposed model to express rights-related exercising activities presented in Table 4.9. The presented work demonstrates that the developed model satisfies the requirements identified in Section 4.4.1 by answering competency questions regarding the modelling of the existence of data subject rights, how and where such rights can be exercised, what data is necessary to fulfil such rights and which entities are in charge of implementing and exercising it, how to keep records of said exercising

activities, including timestamps, the status of the right request activity and other provenance metadata, which rights are applicable according to the legal basis used by the data controllers, and which justifications can be provided by them to not fulfil, delay or exercise a request.

4.6 Alignment with the ISO/IEC 27560 standard

ISO and IEC are an international standardisation body with technical committees established to specify requirements and guidelines for particular technical activities. The ISO/IEC 27560 standard on 'Privacy technologies – Consent record information structure' *"specifies an interoperable, open and extensible information structure for recording PII principals' consent to PII processing"* [ISO/IEC JTC 1/SC 27, 2023]. It outlines requirements and recommendations regarding the utilisation of consent receipts and consent records related to the processing of personally identifiable information (PII). Its objectives include facilitating (i) a consent record for data controllers, (ii) the exchange of consent details among information systems, and (iii) the management of the consent life cycle. Figure 4.13 showcases the information elements of the consent record and receipt structure specified in the ISO/IEC 27560 standard, including examples of what information should be entered in each field.

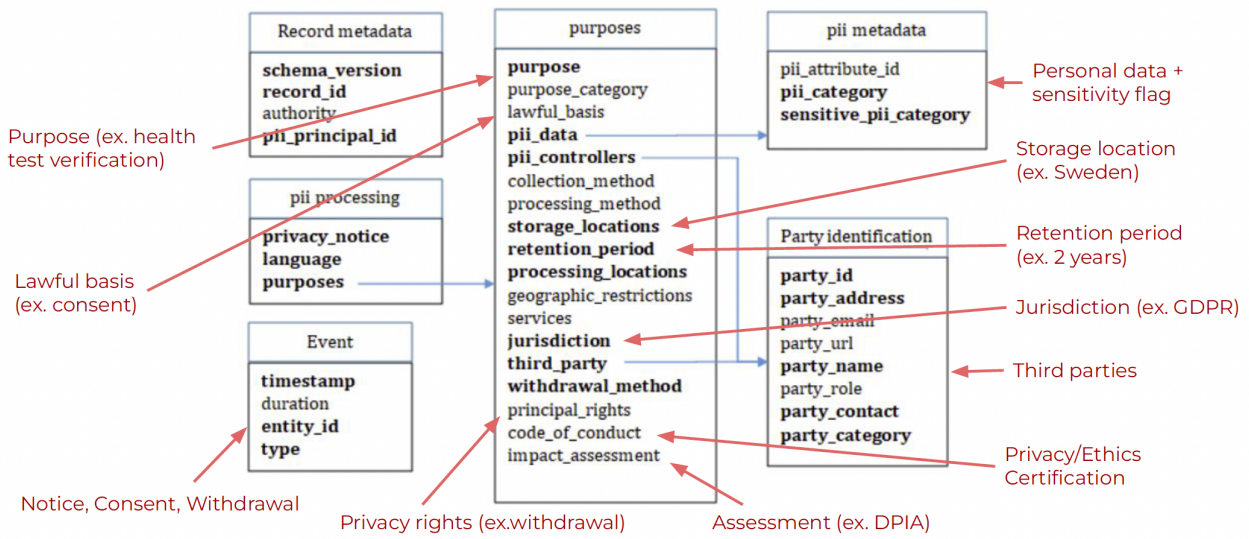


Figure 4.13: Elements of the ISO/IEC 27560 consent record and receipt structure [ISO/IEC JTC 1/SC 27, 2023].

Most of the fields illustrated in this Figure can be represented using the OAC-based data agreements proposed in Section 4.2, as well as the existing ODRL and DPV specifications and the proposed work on PLASMA and the rights exercise extension described in Section 4.3 and 4.4. The 'record metadata' elements **record_id**, **pii_principal_id**, and **authority** can be instantiated using the `odrl:uid`, the `dpv:hasDataSubject`, and the `dpv:hasAuthority` properties, and the 'pii processing' terms **privacy_notice**, and **language** with PLASMA's notice terms and `odrl:language` left operand. Furthermore, the 'Event' terms can be modelled using PLASMA and DPV's concepts to model notices, consent statuses, and logs of processing activities, and records of right exercise activities with the work proposed in Section 4.4. The 'Party identification' elements

Table 4.15: Validation of the competency questions of the proposed model to express rights-related exercising activities with SPARQL queries.

CQR*	SPARQL query
CQR1	`SELECT ?pdh ?right WHERE { ?pdh a dpv:PersonalDataHandling . ?pdh dpv:hasRight ?right . }`
CQR2	`SELECT ?right ?exercise_point WHERE { ?right a dpv:DataSubjectRight . ?right dpv:isExercisedAt ?notice . ?notice a dpv:RightExerciseNotice . ?notice foaf:page ?exercise_point . }`
CQR3	`SELECT ?right ?notice WHERE { ?right a dpv:DataSubjectRight . ?right dpv:isExercisedAt ?notice . ?notice a dpv:RightExerciseNotice . }`
CQR4	`SELECT ?right ?necessary_data WHERE { ?right a dpv:DataSubjectRight . ?right dpv:isExercisedAt ?notice . ?notice a dpv:RightExerciseNotice . ?notice dpv:hasPersonalDataHandling ?necessary_data . }`
CQR5	`SELECT ?right ?implementer WHERE { ?right a dpv:DataSubjectRight . ?right dpv:isExercisedAt ?notice . ?notice a dpv:RightExerciseNotice . ?notice dpv:isImplementedByEntity ?implementer . }`
CQR6	`SELECT ?activity ?data_subject WHERE { ?activity a dpv:RightExerciseActivity . ?activity dpv:hasStatus dpv:RequestInitiated . ?activity dpv:hasDataSubject ?data_subject . }`
CQR7	`SELECT ?activity ?date WHERE { ?activity a dpv:RightExerciseActivity . ?activity dcterms:date ?date . }`
CQR8	`SELECT ?activity ?status WHERE { ?activity a dpv:RightExerciseActivity . ?activity dpv:hasStatus ?status . }`
CQR9	`SELECT ?pdh ?legal_basis ?right WHERE { ?pdh a dpv:PersonalDataHandling . ?pdh dpv:hasRight ?right . ?pdh dpv:hasLegalBasis ?legal_basis . }`
CQR10	`SELECT ?activity ?data_subject ?status ?date ?controller WHERE { ?activity a dpv:RightExerciseActivity . ?activity dpv:hasStatus ?status . ?activity dpv:hasDataSubject ?data_subject . ?activity dcterms:date ?date . ?activity prov:wasAssociatedWith ?controller . }`
CQR11	`SELECT ?activity ?justification WHERE { ?activity a dpv:RightExerciseActivity . ?activity dpv:hasStatus dpv:RequestRejected . ?activity dpv:hasJustification ?justification . }`

can be modelled using the `oac:Entity` placeholder that should be set on agreement policies for assigners and assignees, as well as for constraints on which recipients can get access to the data, which can then be specifically defined with DPV's legal entity concepts as well as PLASMA entity terms to specify specific decentralised/Solid-related roles. Contact-related information can also be modelled with DPV's `hasName`, `hasContact` and `hasAddress` properties. Moreover, most 'purposes' terms can already be modelled with existing and proposed work, such as the proposed OAC terms to restrict purposes, legal bases, data types, processing operations, including collection, or services, as well as existing ODRL temporal and spatial constraints, and DPV's concepts related rights, jurisdictions or impact assessments. Additionally, DPV, and in particular its personal data extension, can be used to specify the categories of personal or sensitive personal data being used and specified in the standard as 'pii metadata'.

As such, it is possible to check that the existing and proposed work is aligned with the ISO/IEC 27560 standard for consent records and hence can be used to fulfil almost all of the recommended elements. Future work can be focused on providing machine-readable codes of conduct, and more comprehensive event specifications, e.g., data holders' permission terms proposed in the DGA.

Chapter 5

Legal and Ethical Challenges of Decentralised Data Environments

The content of this Chapter has already been partially included in the articles published during this Thesis [Esteves et al., 2022a, Asgarinia et al., 2023, Florea and Esteves, 2023].

This Chapter discusses the legal and ethical challenges of the impact of data-driven innovation in society, in particular, related to the emergence of PIMS as a service that helps individuals have more control over the processing of their data. While some studies have recently been published on the intersection of Solid and data protection requirements, as reviewed in Section 2.1.2, plenty still has to be overcome to have a 'legally-aligned' personal datastore. This interdisciplinary discussion relies on the collaborations fostered through the PROTECT project, and other EU-funded projects described in Section 1.5, as well as through the participation in the W3C DPVCG work with data protection law experts.

Section 5.1 describes the emergence of decentralised personal information management systems as a way to give users more control over their personal data and the challenges that still need to be overcome in order to have to a GDPR-aligned personal datastore.

Section 5.2 discusses the usage of OAC policies as a precursor of consent for Solid, which can enable compliance with several GDPR requirements including the transparent information obligations of Articles 13 and 14 and the conditions to obtain valid consent pursuant to Articles 4.11 and 7.

Section 5.3 argues whether the automation of consent can be performed while maintaining the 'informed', 'freely given', 'specific', and 'unambiguous' character of GDPR consent. In particular, the specificity of purposes and processing operations, the distinction between data controllers and recipients, the compatibility of purposes, and the delegation of consent are further analysed through a 'legal+tech' approach, relying on GDPR's requirements and on the OAC and PLASMA implementations.

Section 5.4 discusses the special requirements of GDPR's special categories of data and research-related exceptions and, in particular, the requirements related to the sharing of health data for biomedical research or for the management of public health.

To conclude, Section 5.5 debates the ethical challenges of controlling data and reclaiming control over it and explores how decentralised PIMS can help build confidence in data exchange practices and trust in the providers and developers of such systems.

5.1 The emergence of decentralised PIMS

As previously mentioned in Section 1.3.4, the governance of data flows, and in particular of *personal* data flows, has been a topic of discussion since the early 1970s and 1980s, when the Fair Information Practice Principles (FIPPs) [Cate, 2006] and Convention 108 [Council of Europe, 1981] were first created, to GDPR and subsequent personal data-related regulations being developed in countries such as Brazil or India [Bradford, 2019]. Most of these instruments rely on the existence of an accountable entity that is responsible for establishing the purpose of processing personal data from a natural person, who has rights that must be respected for said processing to be considered compliant with the law. This model has been the most prevalent since most personal data are stored in large centralised databases under the control of only a certain number of Big tech companies, however, it does not account for cases where the processing is shared among different entities which have distinct purposes or rely on unsuitable legal bases, or the information overload that prevents individuals from actually understanding what they are consenting to [Ben-Shahar and Schneider, 2014]. As such, new data governance systems that assist individuals in having more understanding and control over their data and trust in data processing services, such as *data cooperatives*, *data trusts*, *data commons* or *personal data sovereignty* schemes, are being proposed [Viljoen, 2021, Craglia et al., 2021]. For instance, data trusts may give more emphasis on ensuring that data subjects have a good understanding of the purposes for which the personal data is being used and that these are explained clearly and transparently, however, they do not offer control over those purposes in the same way data cooperatives do, where data subjects can participate in purpose and rule decision making. Moreover, these systems are even starting to be regulated, such as the new requirements on data intermediation services described in the DGA [2022g].

In this context, the emergence of decentralised PIMS for the Web, such as the personal datastores model promoted by Solid and studied in this Thesis, has earned many advocates in the last years. In particular, when it comes to trust, the usefulness and ease of use of digital personal datastores have been proven to be an important factor in increasing citizens' trust in personal data-handling services by allowing them to share their sensitive data for the 'public good' while maintaining a sense of control over their data [Mariani et al., 2021]. Moreover, while these decentralised solutions are not without their faults, as has been shown by blockchain-related scandals in the financial services industry [Zetzsche et al., 2019], their Semantic Web-based counterparts have been gaining a large number of adopters recently as such systems can actually allow its users to choose who can access their data and, therefore, actually shift the power balance in favour of the individuals. By detaching the storage of data from the data processing services and promoting the usage of Web standards, individuals can move their data between storage providers, use the same data across different services and choose which services and applications best suit their preferences and needs without being locked out of the access to their data [Verbrugge et al., 2021, Ilves and Osimo, 2019]. This user-managed access to data represents a considerable change from the current *status quo*, where individuals must usually accept an application's privacy policy in order to use it, while personal datastores present the next step towards having an actual negotiation of

privacy terms between individuals and data processing entities. Such systems are also promoted by the EDPS as a mechanism to enable personal data sovereignty where "Individuals, service providers and applications would need to authenticate to access a personal storage centre" in an interoperable manner [European Data Protection Supervisor, 2021]. Additionally, the European data spaces initiative launched by the European Commission [2020] follows the same spirit by encouraging the development of infrastructures for data holders and data users to share and reuse data across different services while respecting European data protection law. However, it should be acknowledged that such data spaces are still focused on encouraging industrial data sharing, for large institutions to gain economic and societal benefits, often at a sectorial level, rather than empowering data subjects individually. Nevertheless, they are still an improvement over current systems where the data controller is purely motivated by profit for their organisation alone.

While personal datastores' developers have as their main banner that data subjects are 'controllers' of their data, this view is incompatible with most data protection-related regulations as *"most [...] legal systems are structured around the identification of an accountable entity"* [Chomczyk Penedo, 2021] which is given duties in order to ensure that their data processing activities do not affect data subjects' fundamental rights. In addition, personal data processing activities often involve a complex web of parties that share control of the usage, storage and collection of data for distinct and shared purposes, a fact that makes the compliance with the information requirements described in GDPR's Articles 12 to 14 quite challenging and difficult to implement [Lovato et al., 2023] – compliance with such requirements is usually dealt with by providing lengthy and complex notices which are not easy to understand and place a significant burden on data subjects as they have to deal with at least one notice for every personal data processing service they use [Terpstra et al., 2019, Linden et al., 2020]. Although these notices usually address the information required by the law, in no way do they fulfil the 'informed' character of consent, as prescribed in GDPR's Article 4.11, as it is utterly challenging for data subjects to understand the plethora of terms and conditions of all the personal data handling services that are used nowadays, from smartphone applications to personalised streaming of content, and to give consent in a freely, specific, informed and unambiguous way [Mohan et al., 2019]. Beyond consent management, notices are also an inefficient way for data subjects to exercise their rights as they are limited to describing rights and where to exercising them, without focusing on providing data subjects with actually tools *"for facilitating the exercise of the data subject's rights [...], including mechanisms to request and, if applicable, obtain, free of charge, in particular, access to and rectification or erasure of personal data and the exercise of the right to object"*, as mentioned in GDPR's Recital 59 [2016b].

As such, recently, there has been legal work in identifying the different roles and responsibilities that distinct entities occupy in decentralised systems and how said systems can be used to facilitate the exercise of data subjects' rights, fulfil the data protection principles of privacy by design and by default and improve the clarity and transparency of personal data handling processes in contrast to the existing landscape [Janssen et al., 2020], as described in Section 2.1.2. Nevertheless, work still needs to be done to align such decentralised systems with the legal requirements, in particular, related to the identification and enforcement of a lawful basis and transparent purpose that justify the access to data. Particularly, in this Thesis, the focus is positioned on how to obtain valid consent in Solid, according to the GDPR, while promoting the usage of automation to improve the current information overload felt by data subjects in terms of consent management. Accordingly, in this Chapter, the introduction of a semantic policy layer, based on the vocabularies described in

Chapter 4, for providing the necessary information to obtain informed and valid GDPR consent is studied from a technical and legal angle.

It is likewise important to distinguish between what can be technologically or legally enforced. Although technically a certain app or service can be restricted to only access certain data, by being allowed to read data from a personal datastore, it can also copy it, even if within a decentralised setting this is not necessary at all. Thereupon, the realm of law comes into play. Although the wishes of the data subjects, as stated by the preferences they have stored in their personal datastore, have a role to play in the negotiation of privacy terms between data subjects and data controllers [Verborgh, 2017], their legal value is still up to debate, as these are quite new technological tools which are still to be argued and tested in the court of law.

As such, in the next Section, the usage of policies as a precursor of consent is further studied.

5.2 Policies as a precursor of consent

This Section discusses the usage of OAC policies as a tool to express consent in advance for Solid and how such policies can enable compliance with several GDPR requirements including the transparent information obligations of Articles 13 and 14. As such, these policies come as a solution to overcome the shortcomings of Solid's access control mechanism when it comes to dealing with GDPR's information requirements. Moreover, by enabling the communication of this information, policies can be used as a tool to fulfil the conditions to obtain valid consent under Articles 4.11 and 7 of the GDPR.

5.2.1 Distinguishing consent from access control

It is important to make a distinction between the legal notion of giving consent and the technical means used to grant an app, service or user access to a resource stored in a decentralised personal datastore such as a Solid Pod.

As previously discussed in Section 2.1.1, Solid Pods are decentralised, permission-based data storage environments, by default. This means that in the absence of a tangible authorisation, resources cannot be accessed by apps or users. Authorisations can then be provided directly and indirectly by accepting requests from apps as they are being received or by setting the rules of access in advance, respectively.

From GDPR's viewpoint, user authorisation is not always required for the processing of personal data, but it also might not be enough for entities to process personal data in a lawful manner in such decentralised settings. In the first case, it might be *unnecessary* as there are other legal bases in GDPR's Article 6.1 which can be used, beyond consent, that do not involve an active choice being made by the data subject, such as the performance of a contract –Article 6.1(b)– or the legitimate interests of the data controller –Article 6.1(f)– [Kranenborg, 2014]. Taking the former as an example, there is no need to have the consent of the data subjects to access personal data when they have entered into a contract with the data controller and access to said data is necessary for the performance of said contract [European Data Protection Board, 2019]. Moreover, if indeed the access is based on the consent of the data subject, then the current status quo of access control in

Solid –whether being the WAC or the ACP authorisation mechanisms– is not enough for obtaining valid consent according to the GDPR, as in Article 4.11 valid consent is described as being a *"freely given, specific, informed and unambiguous indication of the data subject's wishes"* [2016b].

By comparing both the legal and the technical requirements, described in the previous paragraphs, it is possible to arrive at two sets of problematic cases:

(i) instances when app providers have a valid legal basis beyond consent to have access to the data, but do not have access to said data as no permission-based authorisation, granted by the data subject, is stored in the Pod; and

(ii) instances when app providers use consent as a ground for lawfulness, however, the authorisation available on the Pod does not fulfil the conditions for valid consent according to Articles 4.11 and 7.

In this Thesis, the second cluster of cases is explored by discussing whether the introduction of fine-grained access control policies, modelled with OAC, is enough for obtaining valid consent.

5.2.2 Introducing OAC policies in the Solid ecosystem

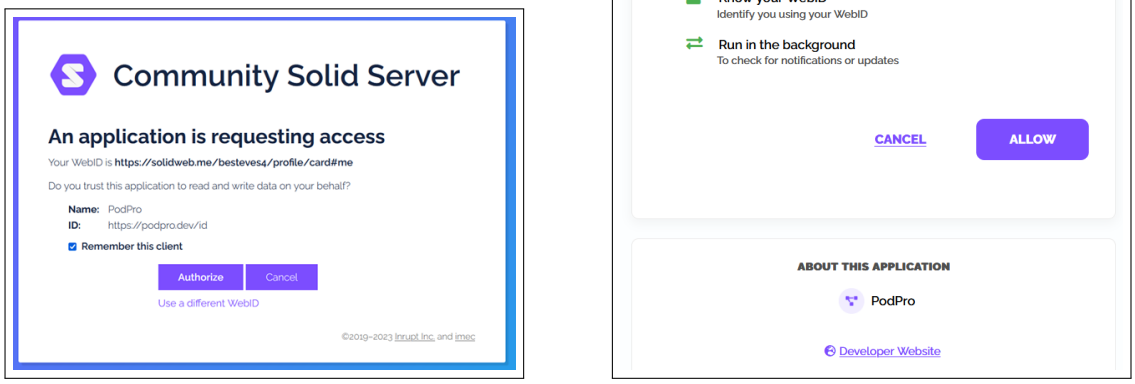
In addition to a lawful basis for processing, Article 5.1(a) states that personal data should be *"processed [...] in a transparent manner in relation to the data subject"* [2016b]. The information obligations described in Articles 12 to 14 depict the required information that data subjects must be provided with, regardless of the chosen legal basis, in order to have transparent information regarding the processing of their personal data. This means that data subjects always have the right to have access to this information, while data controllers are always obliged to provide it, even if the legal basis for processing personal data is not consent. Additionally, Recital 59 provides that *"Modalities should be provided for facilitating the exercise of the data subject's rights [...] free of charge [...]"*. In this context, OAC policies can serve as a modality that enables the data subjects' right to information regarding the processing of their personal data. In particular, OAC-based agreements stored on Solid Pods, such as the one depicted in Listing 4.4, resulting from the matching of user offers and data requests, described in detail in Chapter 6, are thus accessible to data subjects and can be used by them to easily understand whether the specific conditions for accessing data, set on the agreement, vary from their personal preferences stored in the Pod in the form of OAC-based preferences and requirements.

The specific privacy terms that need to be provided by data controllers are specified in GDPR's Articles 13 and 14 and detailed in Table 2.2 as informational items I1 to I19. As can be checked in Figure 4.3 and Table 4.5, the OAC profile provides concepts to express personal data types, legal basis, recipients, purposes for processing, processing operations and the identity of the data controllers accessing the data. This leaves out some elements noted in Article 13.1, namely the controller's contact details and its representative, the DPO's contact details, the legitimate interests of the data controller or third party recipient, if the used legal basis is grounded on Article 6.1(f), and information about the transfer of data to third countries or international organisations. Moreover, Article 13.2, in order to ensure fair and transparent processing, also states that data subjects should be informed about the retention period of the data, the existence of data subject rights, statutory or contractual obligation details if the provision of data is a requirement to enter

into a contract or a statutory obligation, including the possible consequences of failing to provide such data, and the existence of automated decision-making.

If the user's policies do not include all these necessary elements, or if there are discrepancies between them and the controller's access request, then the data subject should be notified about this information at the time of the data request. In Solid, as previously stated in Section 5.2.1, access can be granted by (i) accepting requests when starting to use a new application or by (ii) setting the access rules in advance. In the first case, this is done through an authorisation dialogue, such as the examples provided in Figure 5.1, and, in the second, through a Pod management app, such as Inrupt's PodBrowser[1] in Figure 5.2 or Penny[2] in Figure 5.3. Figure 5.2(a) illustrates the authorisation dialogue related to the Community Solid Server (CSS)[3] and Figure 5.2(b) the Inrupt's Enterprise Solid Server (ESS)[4] Pod and identity providers. While ESS's dialogue includes some information on the purposes for access and CSS's on the specific types of access being provided, as is visible through the Figures, both dialogues do not include all the elements previously discussed for the user to be able to provide informed consent.

Figure 5.1: Screenshot of the authorisation dialogue of the



(a) Community Solid Server (b) Enterprise Solid Server

Furthermore, two more legal challenges should be considered regarding the information obligations set out in the GDPR. The first relates to how the information is presented to the data subject as GDPR Article 12 states that data controllers have an obligation to provide this information *"in a concise, transparent, intelligible and easily accessible form, using clear and plain language"* [2016b].

[1] https://docs.inrupt.com/user-interface/podbrowser/ (accessed on 21 December 2023)
[2] https://penny.vincenttunru.com/ (accessed on 21 December 2023)
[3] https://communitysolidserver.github.io/CommunitySolidServer/7.x/ (accessed on 21 December 2023)
[4] https://www.inrupt.com/products/enterprise-solid-server (accessed on 21 December 2023)

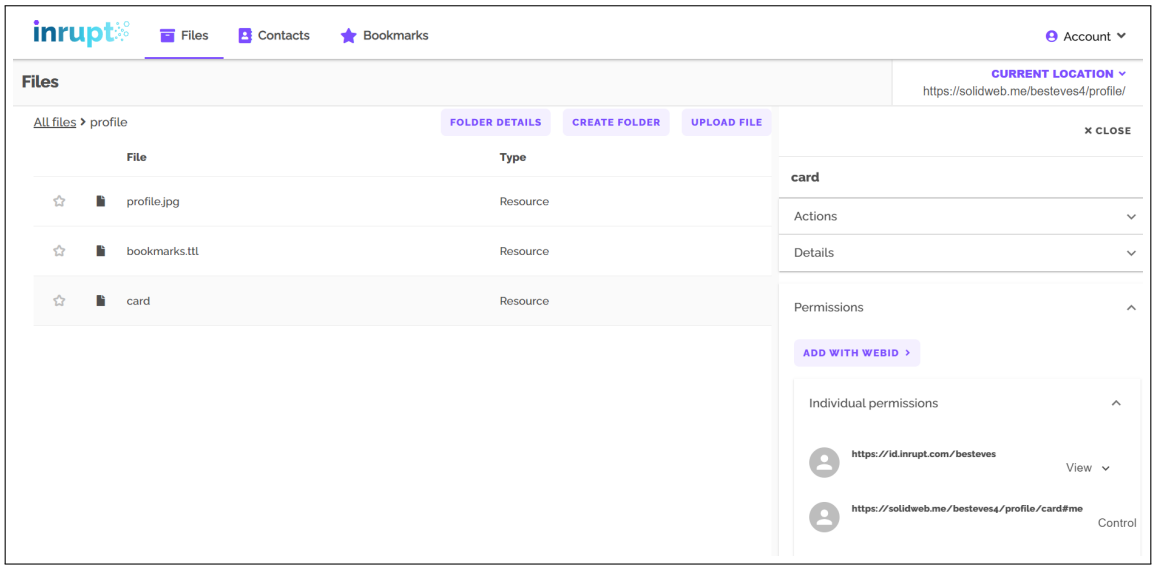


Figure 5.2: Screenshot of Inrupt's PodBrowser app to manage data and access grants.

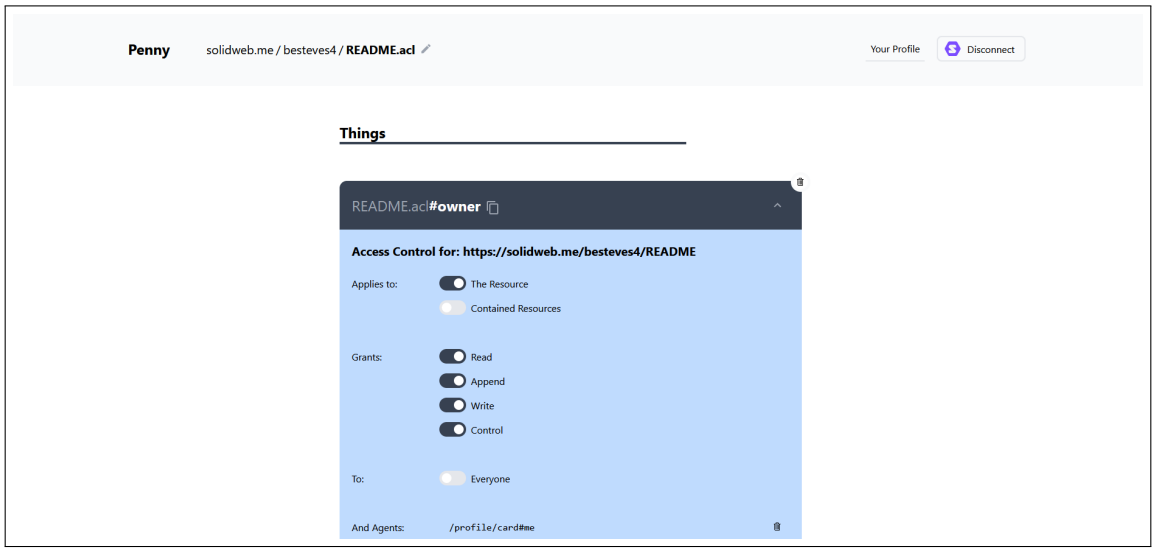


Figure 5.3: Screenshot of Penny app to manage data and access grants.

As such, while the user's policies, others' requests and data access agreements can be easily accessed by data subjects if stored in Solid Pods, the implementation of interfaces to display the result of the policy matching process, especially the information that was previously unknown by the subject, might also be necessary to fully fulfil the requirements of Article 12 in ensuring that data subjects have read and understood this information.

The second challenge is related to the timing of the notification, as Articles 13 and 14 [2016b] set different rules which depend on whether the data collection is done directly from the data subject or another entity. As the GDPR does not directly mention data intermediary services, there is a gap that should be further explored to understand which Article applies in the Solid context. On one hand, if the Pod provider is deemed a data controller, then the personal data is not directly collected from the data subject [Pandit, 2023] and Article 14 applies, meaning that the data subject must be informed *"at the latest at the time of the first communication to that data subject"* [2016b]. Access requests to Solid Pods can be considered to be communications with data subjects, and as such, at the time of the request, the information requirements should be fulfilled. On the other hand, if the Pod provider is not thought to be a data controller, i.e., it is simply considered a piece of software used by the data subject, then the data is directly captured from the data subject and Article 13's requirements [2016b] must be fulfilled at the time when said data is obtained. Regardless, in both interpretations, the data subjects must be notified at the latest in the instant when the requests reach the data subjects' Pod.

Additionally, the informed character of consent is only one of a series of requirements that must be met in order to obtain valid consent. After being informed, data subjects must state their preferences *"by a statement or by a clear affirmative action"* [2016b] which signifies their agreement with the handling of their personal data. Moreover, EDPB's and WP 29's guidelines on consent [European Data Protection Board, 2020a, Article 29 Data Protection Working Party, 2011, 2016] further develop the freely given, specific, informed and unambiguous characters of consent. Among the discussed topics, these entities' guidelines state that consent must be granular, the data subjects must be aware of the consequences of refusing to consent and the distinct purposes for processing data must not be tied together. Thus, simply accepting an access request does not necessarily signify consent according to the GDPR.

5.2.3 Expressing consent in advance through OAC policies

The GDPR does not forbid the expression of consent in advance. In fact, Recital 32 mentions that *"Consent should be given by a clear affirmative act [...], such as by a written statement, including by electronic means, [...]. This could include ticking a box when visiting an internet website, choosing technical settings for information society services or another statement or conduct which clearly indicates in this context the data subject's acceptance of the proposed processing of his or her personal data"* [2016b]. Nevertheless, in order for consent to be valid under the GDPR jurisdiction, it must be specific even for circumstances that may not have already happened [Kosta, 2013] and the data controllers must be able to demonstrate that *"the data subject has, by active behaviour, given his or her consent to the processing of his or her personal data and that he or she has obtained, beforehand, information relating to all the circumstances surrounding that processing, in an intelligible and easily accessible form, using clear and plain language, allowing that person easily to understand the consequences of that consent, so that it is given with full knowledge of the facts"*, as stated in Case

C-61/19 held in 2020 at the European Court of Justice [2020]. As such, in this Section, the usage of OAC policies to express the required GDPR terms to have valid consent is further explored.

The automation of consent on the Web is not a new idea. As a matter of fact, the Do Not Track (DNT)[5] initiative and the previously described P3P are two examples in this respect. Although none of these solutions have succeeded in being consumed at a large scale, they can be illustrative use cases of what to do – and do not do – while developing a system for Web consenting. The DNT initiative focused on blocking the ad-tech industry from tracking users based on their online behaviour by sending a signal from the users' browser to all Web pages they visited with the preference to not be tracked, similarly to the right to object asserted in GDPR Article 21, however, it failed due to the lack of browser adoption and enforcement mechanisms [Kamara and Kosta, 2016]. Similarly, the P3P initiative allowed Web pages to *"express their privacy practices in a standard format that can be retrieved automatically and interpreted easily by user agents"* and *"enable an expanded ecosystem in which web sites would consistently inform web user agents of personal data collection intentions and web users would configure their individual user agents to accept some practices automatically [...]"* [Cranor et al., 2002b]. However, one of P3P's main drawbacks was the lack of consistency between human and machine-readable privacy notices communicated to users [Cranor, 2002], a challenge which can also be attributed to Solid as the information presented to users in consent dialogues is not aligned with the authorisation statements stored in Solid Pods. Moreover, WP 29 also stated that P3P had the capability of misleading data controllers into believing that they could be discharged of certain obligations as long as data subjects had already agreed to the processing of their data [Article 29 Data Protection Working Party, 2014], an issue which can also very easily be relevant for the Solid ecosystem.

Nevertheless, Solid differs from P3P in the sense that it provides its users with a decentralised storage unit equipped with a permission-based access control mechanism, i.e., access to data is only provided in the presence of an authorisation for a particular application or user. Furthermore, in such decentralised systems there is no need to transfer or make copies of data as access can be provided on demand to any user or application through its authorisation and authentication mechanisms, removing the need for such entities to keep copies of data in their own servers. Said mechanisms can also serve as the starting point to keep access and usage logs in Solid Pods, which can be used by users and by external auditing entities to check whether Web services are using data according to their announced policies. Such transparency logs are also of the utmost importance in cases where the controller provides a valid justification to have a copy of the personal data. As such, users will have a more transparent overview of how their data is being used, which comes as an improvement over P3P's lack of consistency and policy enforcement – *"no enforcement action followed when a site's policy expressed in P3P failed to reflect their actual privacy practices"* [Cranor et al., 2002b] –, the main issues that led to its failure. What's more, as previously stated in Section 2.1.2, there is ongoing research on the modelling of usage control policies [Akaichi et al., 2023] and enforcement mechanisms for Solid [Slabbinck et al., 2023].

The usage of pre-configured choices is also discussed in Recital 66 of Directive 2009/136/EC, the successor of the ePrivacy Directive [2002], which states that *"Third parties may wish to store information on the equipment of a user, or gain access to information already stored, for a number of purposes, ranging from the legitimate (such as certain types of cookies) to those involving unwarranted*

[5] https://www.eff.org/issues/do-not-track (accessed on 28 January 2024)

intrusion into the private sphere (such as spyware or viruses). [...] Where it is technically possible and effective, [...] the user's consent to processing may be expressed by using the appropriate settings of a browser or other application" [2009]. This is also reflected on a few national implementations of the ePrivacy Directive that allow the indication of consent via technical means, e.g., Romania's Legea 506/2004 states that *"the subscriber or user can use the settings of the internet browsing application or other similar technologies to delete stored information or to deny access to such information to third parties"* [2004]. However, the utility of browser settings to express consent is being challenged as recently the data protection authority in Finland ruled that *"instructing website users to accept or decline to the use of cookies through browser settings does not constitute active and explicit consent under the GDPR"* [Fich, 2021]. Nonetheless, several technical solutions to signal users' preferences have been emerging recently, e.g., the Global Privacy Control (GPC) [2021] or the Advanced Data Protection Control (ADPC) [Human et al., 2021] *'privacy signals'*, which are still lacking adoption due to a lack of standardisation and legal approval towards the fulfilment of ePrivacy requirements [Santos and Pandit, 2023].

In the next Section, the building blocks for consent automation are explored in detail, with a reflection on how OAC and Solid can be adapted to fulfil legal requirements.

5.3 Can consent be automated?

To discuss consent automation, first one should look into the rationale of why it is such an important requirement in personal data protection law. As per Jarovsky [2018], the main rationale behind consent is to retain human autonomy and to enable data subjects to have agency regarding the processing of their data. To achieve that, data subjects must (i) comprehend the circumstances that surround the processing of their data, (ii) decide which is their optimal choice among a variety of options, and (iii) express their choice, while knowing that they can change it at any point in the future. However, if there are no technical and/or organisational measures in place to preserve individual autonomy in this process, meaningful, freely given consent cannot be achieved due to *"issues of cognitive limitations, information overload, information insufficiency, lack of intervenability and lack of free choice"* [Jarovsky, 2018]. Solove [2012] also outlined a few shortcomings in the self-management of privacy, distinguishing between cognitive limitations related to human decision-making abilities and structural limitations that prevent an adequate cost-benefit analysis of consenting to simultaneous personal data processing activities. As such, presenting Solid users with a consent dialogue with the result of the matching for each access request that comes in will result in similar scalability issues for the users [McDonald and Cranor, 2008].

One possible solution to the previously identified consent-related issues is to automate some aspects of giving consent [Baarslag et al., 2017]. However, this solution has often been criticised due to the complexity surrounding current personal data processing activities on the Web, which might compromise the validity of consent [Jarovsky, 2018, Solove, 2023]. Consenting is context-dependent and encompasses weighing the risks, likelihood of harms and benefits of several variables involved in a personal data processing activity. As such, it is difficult to imagine how an automated system can weigh all the arguments in favour and against the processing of personal data, while maintaining the interests and autonomy of the data subject at the center of the decision making

algorithm.

Thus, in this Section, the setting of OAC user policies in advance and the matching of such policies with requests for data access is analysed to check if such a system is sufficient to comply with the legal requirements for expressing valid consent. Consenting is usually a binary choice – the data subject either agrees with the conditions set by the data controller to process their data or they do not. Nevertheless, different privacy laws implement this choice in distinct manners: in the United States the 'opt-out' choice predominates, i.e., *"organizations post a notice of their privacy practices and people are deemed to consent if they continue to do business with the organization or fail to opt out"*, while in the EU the 'opt-in' option prevails, i.e., *"people must voluntarily and affirmatively consent"* [Solove, 2023]. As such, the latter involves (i) the data controller requesting consent and (ii) the data subject accepting or rejecting it. If users set their policies in advance, this order is inverted. Even though the GDPR does not regulate the interaction between data subjects and software to assist them in expressing consent, as previously mentioned, Recital 32 [2016b] suggests the usage of *"technical settings"* to indicate *"acceptance of the proposed processing of his or her personal data"*.

Moreover, two levels of automation, both triggered by a data request, can be considered: (i) the result of the matching, between user policies and data request, is presented to the user for him/her to consent, or (ii) access to data is given automatically if the data request matches with the user's policies. The former – the consent dialogue, based on the policy matching algorithm – improves the transparency of the processing activity and helps the data subject to make an informed choice, while the latter assists with the issues related to information overload and scalability.

5.3.1 Expressing specific consent

In this Section, the specific character of consent is going to be analysed to understand whether OAC can be used to automate access to data in Solid Pods in a GDPR-aligned manner. According to Article 4.11 [2016b], consent must express a specific *"indication of the data subject's wishes"* that *"signifies agreement to the processing of personal data relating to him or her"*. However, the wording "indication of wishes" is rather vague in the sense that such wishes might be related to the categories of personal data, the purpose for processing, the processing operations, the identity of data controller(s) and/or third party recipients, or their interconnection. Moreover, the European Court of Justice mentions in Case C-61/19 Orange Romania [2020] states that the data subject's wishes *"must relate specifically to the processing of the data in question and cannot be inferred from an indication of the data subject's wishes for other purposes"*. The EDPB also included guidance on the specificity of consent in its consent guidelines [European Data Protection Board, 2020a], in particular related to (i) using purpose as a safeguard against 'function creep'[6], (ii) the granularity of consent requests, and (iii) the requirement to provide information related to consent separately from other data processing matters. Moreover, pursuant to Recital 42 [2016b], for consent to be informed, the data subject should be aware of, at least, the purpose for the personal data processing and the identity of the controller(s). However, the level of detail in which the purpose must be described is not further prescribed in the regulation. According to Kosta [2013], the specificity of

[6] Koops [2021] defines 'function creep' as *"an imperceptibly transformative and therewith contestable change in a data-processing system's proper activity"* or, in simpler terms, *"the expansion of a system or technology beyond its original purposes"*.

consent is fulfilled when the relation between personal data and its processing, as well as all other conditions surrounding the processing activities, are explained. Furthermore, consenting should be as specific as needed for safeguarding the data subject's right to informational self-determination[7]. As such, the following analysis will be focused on the purpose of processing personal data, as well as on the identity of the data controller, and how they relate to the specific character of consent.

Distinguishing the processing operation from the purpose for processing The GDPR explicitly mentions that not only the purpose but also the processing operation needs to be specific to have valid consent. Article 6.1(a) [2016b] states that *"Processing shall be lawful only if [...] the data subject has given consent to the processing of his or her personal data for one or more specific purposes"*, while Recital 43 [2016b] pinpoints that *"Consent is presumed not to be freely given if it does not allow separate consent to be given to different personal data processing operations despite it being appropriate in the individual case"*. Hence it is important to distinguish between both. While processing operations refer to the actions performed over data – personal data in the case of GDPR –, the purpose expresses the motive or objective of the data controller for processing personal data. This also means that several processing operations might be needed to reach a purpose and, on the other hand, distinct purposes can be reached through the same operation, with use of data being a case in point since it has a very broad scope. As such, *"collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction"* are examples of processing operations set out on Article 4.2 [2016b], while Article 5.1(e) [2016b] mentions purposes such as *"archiving purposes in the public interest, scientific or historical research purposes or statistical purposes"*. Moreover, the distinction between both is also made explicit in Recital 32, which states that *"Consent should cover all processing activities carried out for the same purpose or purposes. When the processing has multiple purposes, consent should be given for all of them"*. As such, this provision can be interpreted as follows: (i) if a processing operation is used to reach more than one purpose, then consent must be obtained for each purpose, and (ii) if multiple processing operations are needed to reach a single purpose, then consent must be obtained for each processing operation.

Moreover, WP 29, in its opinion on the definition of consent, affirmed that *"There is a requirement of granularity of the consent with regard to the different elements that constitute the data processing: it can not be held to cover 'all the legitimate purposes' followed by the data controller"* [Article 29 Data Protection Working Party, 2011], i.e., consent should be specific in relation to a purpose. The relation between processing operations and purposes is also further commented on by WP 29 on these guidelines – *"it should be sufficient in principle for data controllers to obtain consent only once for different operations if they fall within the reasonable expectations of the data subject"* [Article 29 Data Protection Working Party, 2011], however, no further guidance is given on what constitutes 'reasonable expectations of the data subject'. These can be identified, for instance through user studies, however, the result of such studies would only be statistically relevant to the average data subject and not to the particular data subject who has to give consent. Additionally, the contextual integrity theory of privacy developed by Helen Nissenbaum [Nissenbaum, 2004] could also be

[7]The right to informational self-determination was first formulated in German law and it has had a profound impact in European data protection law as it asserts that an individual should have the authority *"to decide fundamentally for herself, when and within what limits personal data may be disclosed, [...]"* [Vivarelli, 2020].

applied to determine the contextual nature of the processing operation. Nissenbaum states that privacy should be considered a right that the individual has over the appropriate, context-based flow of their personal information according to context-specific social norms. As such, the context and norms governing the exchange of personal data should be used to calculate whether the data subject's given consent is specific or not. On the other hand, there is no clear guidance on what should be the granularity of the processing operations. Furthermore, these guidelines also provide an example of how consent fails to be specific – data that is collected for the purpose of providing movie recommendations cannot be used to provide targeted advertisements as the former is more specific than the latter.

However, in a later guidance document, WP 29 also mentions that there are no tools to assess the specificity of data processing elements such as the processing operation or the purpose [Article 29 Data Protection Working Party, 2016]. Furthermore, in its opinion on electronic health records [Article 29 Data Protection Working Party, 2007], WP 29 had already advanced that *"'Specific' consent must relate to a well-defined, concrete situation in which the processing of medical data is envisaged. Therefore a 'general agreement' of the data subject e.g. to the collection of his medical data for an EHR and to subsequent transfers of these medical data of the past and of the future to health professionals involved in treatment would not constitute consent"*. This document also reasons that if the purpose for processing changes at some point in time, then the data subject must be notified to re-consent to the new personal data processing activity and provided with information related to the repercussions of rejecting to consent to such changes.

At this point, it can be discussed whether any change in the purpose, however small, means that consent must be given again by the data subject. Perhaps in the case of a minor change, re-consent could be considered unnecessary, however, the criterion to measure such a change is not clearly defined in the law. As such, it is essential to analyse the matching of offers and requests, for access to data stored in Solid Pods, to understand if the specific character of consent is respected. A policy matching algorithm based on OAC, such as the one detailed in Chapter 6, functions on the basis of subsumption between the data requests and the user policies defined in advance. Thus, by hypothesis, user policies can be broader than data requests, which can be used to doubt the specificity of the consent. However, since OAC policies can also include the usage of prohibitions, these can be used to narrow the scope of the data subject consent, making it more specific.

Applying the purpose limitation principle to assess the specific character of consent GDPR's Article 5.1(b) [2016b] specifies the 'purpose limitation' principle which states that personal data should be *"collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes"*. As such, to fulfil this principle, two requirements should be taken into account: (a) the purpose specification requirement and (b) the non-incompatibility requirement [Koning, 2020]. Previously, the WP 29, in its opinion on the purpose limitation [Article 29 Data Protection Working Party, 2013], stated that all contextual information should be taken into consideration to determine the actual purpose of the personal data handling activity, including the *"common understanding and reasonable expectations of the data subjects"*. Furthermore, Article 8 on the 'Protection of personal data' of the Charter of Fundamental Rights of the European Union [2000] also states that *"[s]uch data must be processed fairly for specified purposes and on the basis of the consent of the person concerned or some other legitimate basis laid down by law"*. In essence, with regard to the processing of personal data, the purpose

element relates to the 'why' the processing is/will be occurring, while the 'how' is related to the processing activity operated over the data [Koning, 2020].

The requirement for the purpose of processing to be specific, in the same manner as the requirement for consent to be specific, is connected with control, self-determination, and autonomy [Koning, 2020]. Additionally, to comply with other principles, such as the 'data minimisation' and 'storage limitation' principles, the purpose for processing should also be considered [Koning, 2020]. Moreover, the motivation for consent – and by consequence for the purpose – to be specific is related to avoiding the broadening or fading of the purpose, which can result in the misuse of personal data by data controllers and recipients and loss of control by data subjects [European Data Protection Board, 2020a]. Without control, users fail to preserve their autonomy and cannot exercise agency over the processing of their data [Jarovsky, 2018]. As such, purpose specification also serves as a measure to mitigate power and information asymmetries. In the absence of such a measure, the power balance leans towards the data controller as it can use the data in its possession according to its interests.

Moreover, the EDPB also discusses how the specific character of consent can be used to mitigate the risk of function creep. According to Koops [2021], the 'creep' component is related to the imperceptibility of the change, which deprives the data subject of the chance to oppose the change and assess its consequences. With the proposed OAC-based system, the risk of function creep is mitigated since the purpose is specifically instantiated in the request for data. Furthermore, this information can also be directly consulted by the data subjects as the OAC system proposes to store user policies, data requests, and data agreements in their Pod, enabling them to exercise their rights, e.g., the right to withdraw consent. Finally, when it comes to the granularity of the purpose for processing, the system proposed in this Thesis can also be easily extended to include a domain-specific taxonomy of purposes, e.g., health, medical, and biomedical research purposes.

Identifying the data controller Using OAC, data subjects can specify explicit policies for specific data controllers that are identified at the time these policies are established. Nevertheless, to grant such authorisation in advance, data subjects need to receive information when initially accessing a Solid application or have access to it elsewhere for review, for instance, through metadata available on an app store. Presently, this crucial information is absent from Solid protocol implementations, as depicted in Figure 5.2(a), which displays the current authorisation dialogue presented to Solid users who use the CSS. While the app's name and ID are displayed, no additional information such as contact details, policies, or links to policies, is provided. Furthermore, there is no official Solid app store that includes metadata about the providers and/or developers of these applications.

In this scenario, an access control policy can vary in specificity and in the way it is modelled. One approach is to mandate that the data subject identifies the data controller by providing their name and contact information, in order to allow them access to personal data. This approach does not present issues in terms of specificity, although the drawback is that the data subject would need to consent to each new data controller individually. On the other hand, another option could be to authorise controllers based on specific criteria, such as industry, country of incorporation, or sector. In this approach, the data subject would establish a set of restrictions for data controllers without specifying concrete entities. While this option offers flexibility and allows for automation,

it is questionable whether it can be considered valid consent as the specificity criterion is not respected.

As such, the initial challenge is related to the moment when the data subject becomes aware of the data controllers' identity. In the first approach, the data controller's identity is disclosed beforehand, while in the second, the data subject establishes the criteria that a requester must meet and the specific identity of the data controller only becomes accessible when the policy matching algorithm determines that the requester meets these criteria. This last option allows for automation – the data subject does not explicitly acknowledge or approve the identity and contact details before granting access, but this information is accessible for review within the data subject's Pod. Both options are possible with an OAC-based system, but involve different policy modelling and matching features.

By taking a look at GDPR's Recital 42 [2016b], it is clear that informed consent requires that the data subject is *"aware at least of the identity of the controller and the purposes of the processing for which the personal data are intended"*. What's more, the European Data Protection Board [2020a] specified that if consent is relied upon by multiple controllers, including controllers to which the data was transferred, all should be identified. Comparably, the Article 29 Data Protection Working Party [2018] highlights the importance of disclosing the controller's identity by stating that a change of identity should **always** be communicated to the data subjects, e.g., through a privacy notice, similarly to a change in purpose or how data subjects can exercise their rights. As such, it can be clearly noted that the specific nature of consent and its informed character are closely linked. Consequently, consent is likely invalid if the data subject is unaware of the identity of the entity processing their personal data at the time of consenting.

As outlined in Section 5.2.2, data subjects must receive information about the processing of their data at the point of data collection or during the initial interaction between them and the data controller. If the decision to grant access is automatically made by the matching algorithm, without the involvement of the data subject, this scenario is unlikely to meet the requirements for valid consent as the data subject is not given the opportunity to make a choice. However, if the legal basis is not consent, the result of the matching algorithm could serve as an informational mechanism and automation can be allowed.

Indeed, Recital 39 [2016b], which discusses the lawfulness, fairness, and transparency principle, emphasises that transparency in particular should be reflected in the *"information to the data subjects on the identity of the controller and the purposes of the processing and further information to ensure fair and transparent processing"*. Transparency is also highlighted in Recital 58 [2016b], which emphasises its significance in contexts where the involvement of multiple actors and technological complexity makes it challenging for data subjects to understand who is processing their personal data and under which conditions. As such, both Recitals indicate that disclosing the identity of data controllers is a crucial aspect of fulfilling information and transparency requirements.

Distinguishing the data controller from the recipients of personal data GDPR's Articles 13 and 14 [2016b] outline a set of information items that must be shared with the data subject. Among them, *"the identity and the contact details of the controller and, where applicable, of the controller's representative"* and *"the recipients or categories of recipients of the personal data"* are clearly separated in Articles 13.1(a) and 14.1(a), and Article 13.1(e) and 14.1(e), respectively. This

differentiation implies that, under specific circumstances, recipients may be identified by category rather than by specific identity details. As such, in this Section, the following legal questions are debated: (i) what distinguishes data controllers from recipients, particularly in a decentralized environment, (ii) are requesters considered data controllers or recipients, and (iii) what is the significance of this distinction.

According to the GDPR (Article 4.9 [2016b]), a recipient is *"a natural or legal person [...] to which the personal data are disclosed, whether a third party or not"*, while a third party is, according to Article 4.10 [2016b], *"a natural or legal person, [...] other than the data subject, controller, processor and persons who, under the direct authority of the controller or processor, are authorised to process personal data"*. Moreover, the EDPB's guidelines related to the concepts of controller and processor [European Data Protection Board, 2020b] state that, unlike the defined roles of controller and processor, the GDPR does not establish distinct obligations or responsibilities for recipients and third parties. These roles are considered relative concepts, describing a relationship with a controller or processor from a particular standpoint. For instance, when a controller or processor shares data with another entity, regardless of the recipient's role as a controller or processor, they are classified as a recipient. Previously, the WP 29 guidance on transparency [Article 29 Data Protection Working Party, 2018] stipulated that controllers must provide information about the recipients or categories of recipients of personal data. Additionally, to uphold the principle of fairness, controllers are obligated to furnish data subjects with information on recipients that is most relevant and meaningful to them, which *"in practice, [...] will generally be the named recipients, so that data subjects know exactly who has their personal data"*. However, WP 29 also acknowledges the option of disclosing the categories of recipients to data subjects. This would require these categories to be identified as specifically as possible, delineating the type of recipient by referencing their activities, industry, sector, sub-sector, and location.

As for the data controllers, due to their crucial role in the provision of data subject rights, should **always** be explicitly identified. Recital 39 [2016b] underlines the importance of transparent communication by stating that individuals should be informed about the risks, rules, safeguards, and rights associated with the processing of their personal data. Hence, the alignment of user policies with actual data requests empowers individuals to assert their rights despite data subjects not expressly acknowledging or agreeing to the identity and contact details of requesters, since this information is accessible in their Pod, enabling them to exercise their data subject rights of access, rectification, erasure, or consent withdrawal as outlined in Articles 15, 16, 17, and 7(3) of the GDPR [2016b], respectively.

Thus, the stringent information criteria in GDPR's Articles 13 and 14 [2016b], requiring the identification of data controllers by name and contact details, appear to be tailored to the current web landscape, where only a handful of entities collect data and subsequently share it with other parties. Despite that, within the ecosystem facilitated by Solid, data exchange occurs directly between the data subject (via the Solid Pod) and an unspecified number of app providers, developers, and/or other users. This poses significant challenges in identifying all these entities by name and contact details, in particular for access automation. In addition, Vogel [2022] examines this issue within the context of the Data Governance Act and contends that it presents a barrier to offering intermediary services as well.

In summary of this Section, it is probable that consenting to the processing of personal data without

explicitly identifying the providers of Solid services, particularly in terms of their identity and contact information, will not meet the criteria for valid consent under the GDPR. Nevertheless, according to certain interpretations and referencing the WP 29 guidance, Pod-stored policies might function as an informational mechanism that facilitates compliance with Articles 13 and 14 [2016b]. Furthermore, through registries of entities, e.g., documented in Solid Pods through the PLASMA vocabulary developed in Section 4.3, information regarding the identity and contact information of both controllers and recipients, as well as other parties, can be kept and consulted by the data subject at any time.

5.3.2 Is already-given consent valid for compatible purposes?

In this Section, the validity of consent is discussed concerning the compatibility of purposes. To give an example, if a research participant indicates a preference to have their data processed for research on Alzheimer's disease, a type of degenerative disease, can this consent also be applied to a request for utilising their data for research on dementia, also a type of degenerative disease?

As previously outlined in this Section, the proposed matching algorithm running over OAC policies currently relies on subsumption. This means that if a user policy grants access for purpose A, and a data request for purpose B (a subclass of A) is made, then access should be allowed. The same principle applies to other matching operations, such as matching on processing operations or categories of personal data. To illustrate it with the example introduced above, if a participant's user policy permits their data to be used for research on Alzheimer's disease but a request is made for research on degenerative diseases, access is denied since the user policy's purpose is more specific than the request. Conversely, if the user policy allows data usage for research on degenerative diseases and a request specifically for Alzheimer's research is made, access is granted as the request's purpose is more specific. Another benefit of using OAC is its capability to express prohibitions, enhancing the specificity of consent. Users can authorise data access for medical research while prohibiting it for specific areas within medical research, such as genetic engineering research.

Nevertheless, OAC currently lacks consideration for the matching of "compatible purposes". For instance, if a participant's user policy permits data usage for Alzheimer's research, and a request is made to utilise the data for dementia research, should access be granted based on compatible purposes? Since dementia, like Alzheimer's, falls under the category of degenerative diseases, it can be argued that access should indeed be permitted. As such, introducing a "compatibility matching" algorithm to Solid would enhance the model, and in particular the automation of access to data, however, it is crucial to assess the legal implications of this addition.

Firstly, it must be discussed how one can determine the compatibility of purposes and whether users desire such a model to complement the policy matching algorithm that governs access to their data. To facilitate this matching, particularly in the context of biomedical research, the example use case developed in Chapter 6 on policies for health data-sharing can be leveraged. This development specifies a taxonomy of health-related research purposes, links to other ontologies containing disease taxonomies, and utilises OAC's matching algorithm. Nonetheless, for the algorithm to assess compatibility, this information must be integrated into the used taxonomies of purposes. This could be achieved, for example, by incorporating a triple statement in the purpose taxonomy

indicating that `:purposeX :isCompatible :purposeY`.

Within GDPR, the concept of purpose compatibility, or the *"further processing [...] shall [...] not be considered to be incompatible with the initial purposes"* non-incompatibility principle, is addressed in the second aspect of the purpose limitation principle outlined in Article 5.1(b). The criteria for evaluating compatibility are further detailed in Article 6.4 [2016b] and quoted below:

(a) *"any link between the purposes for which the personal data have been collected and the purposes of the intended further processing;"*

(b) *"the context in which the personal data have been collected, in particular regarding the relationship between data subjects and the controller;"*

(c) *"the nature of the personal data, in particular, whether special categories of personal data are processed [...];"*

(d) *"the possible consequences of the intended further processing for data subjects;"*

(e) *"the existence of appropriate safeguards, which may include encryption or pseudonymisation."*

From a technical viewpoint, the initial two criteria, namely (a) the association between purposes for use and secondary reuse of data and (b) the processing context, could potentially be evaluated through automated methods, as previously discussed. However, the evaluation required for the third and fourth criteria entails linking the nature of the data with the potential ramifications of its usage for the data subjects. Regarding the final criterion, automated verification of such safeguards can be partially conducted through certifications. Nonetheless, determining the adequacy of such measures presents a challenge for automation, as it requires a risk assessment related to data subjects' rights and interests.

Secondly, regarding the legal hurdles associated with employing these criteria to match user policies with data requests from third parties, the non-compatibility requirement stands as a form of usage limitation. This requirement prohibits the processing of personal data for purposes that are incompatible with the purpose specified at the time of data collection. Moreover, the requirements for compatibility and usage of an appropriate legal basis are conditions that must be met together, i.e., while purpose compatibility is important, it cannot compensate for a lack of legal ground for processing. Therefore, if the purposes are found to be compatible but there is no lawful basis for processing, either renewed consent must be obtained or an alternative legal basis must be identified. In the example used throughout this Section, even if the data subject previously consented to the use of their data for Alzheimer's disease research purposes and the matching algorithm confirmed the compatibility of purposes, this alone would not be sufficient to grant access to the resources in the Pod. Thus, expressing consent for a specific compatible purpose (such as research for dementia) would also be required. Additionally, as will be discussed in the next Section, if special categories of data are being processed, an exception under Article 9 of the GDPR [2016b] must also be identified and recorded in consent notices.

Finally, while it cannot substitute for the lack of consent, the assessment of purpose compatibility is still beneficial since it serves as a prerequisite for reusing data for new purposes under different legal bases.

5.3.3 The intricate boundary between expressing and delegating consent

This Section examines whether an OAC-based system enables data subjects to either express or delegate consent. Moreover, the question of who is liable in the event of errors in such decentralised data-sharing environments is also debated.

The proposed OAC-based matching algorithm converts a preference regarding personal data processing – *"An individual's preferred outcome for a specific privacy-related situation"* – into a decision – *"What an individual chooses to do in a specific privacy-related situation among available options"* [Colnago et al., 2022]. This aligns with the work of Colnago et al. [2022] which primarily focuses on employing these concepts in empirical studies concerning attitudes toward privacy. However, this differentiation could also be valuable in discussing the proposed matching of users policies with data requests within the framework of Solid.

When establishing their preferences through the OAC profile, data subjects delineate specific conditions under which their personal data may be accessed. This also implies that they understand that the practical outcome of the matching is determined through a series of actions adhering to those conditions, including subsumption and exclusion. Thus, consent is extended not only to the categories of data, purposes, and entities acting as data controllers but also to the mechanism that facilitates the transformation of these preferences into choices.

Distinct authors have also explored the delegation of consent from individuals to other entities. Boers et al. [2015] coined the term *"consent for governance"* by arguing that consent, in particular within the realm of biobanking, should be centered on the governance structure of a biobank rather than on the specific details of individual studies. Le Métayer and Monteleone [2009] examined the concept of automated consent from the perspective of shifting from consenting to the use of personal data to consenting to the use of a privacy agent, *"a dedicated software that would work as a surrogate and automatically manage consent on behalf of the data subject"*. Sheehan [2011] explores the notion of ethical consent and distinguishes between first-order and second-order decisions. Second-order decisions are intrinsically different from first-order decisions in that the decision-maker's focus lies on the decision-making process rather than the content of the choice. Sheehan illustrates this with an example involving ordering food in a restaurant. He describes a scenario where a group of friends dine together, and before the waiter takes their orders, one of them briefly steps away and asks another person at the table to place an order on their behalf. When making decisions about delegating decision-making, the individual selects based on factors such as trust in their companions, their knowledge of their taste in food, and the information they possess about the approximate amount they wish to spend.

When applying the differentiation between first- and second-order decisions to OAC, the key inquiry revolves around whether consenting to access conditions and the matching algorithm constitutes a first- or second-order decision. For instance, if a data subject consents to research for the public interest, this concept is further interpreted by the matching algorithm and the ultimate decision to approve a request becomes detailed and specific. However, this decision is not directly made by the data subject but rather indirectly, by delegating it to an OAC-based system. An enhancement suggestion for OAC involves not merely augmenting the ontology with more information, but rather enabling the algorithm to 'learn' to make such inferences. However, in this scenario, consideration has to be given to GDPR's Article 22 [2016b] concerning automated

decision-making.

In cases where consent is delegated rather than directly given by the data subject, it becomes essential to examine the impact of this consent arrangement among the data subject, the agent, and the app provider. It can be argued that the app provider cannot solely rely on the results of the matching algorithm to demonstrate the obtaining of valid consent. From the standpoint of private law, this matter is addressed within the framework of mandate agreements. According to the 'appearance principle', under specific conditions, the intentions expressed by the agent are legally binding over the ones of the data subject. If the choices made by the agent do not accurately represent the data subject's intentions, any discrepancies are resolved between the data subject and the agent [Le Métayer and Monteleone, 2009]. This approach offers legal certainty within contract law. However, the data subject has more protection under European data protection law. According to the GDPR, the app provider, acting as the data controller, is responsible for ensuring and demonstrating that consent was validly obtained (Article 6.1(a) and 7.1 [2016b]). In the context of decentralised systems as Solid, this entails the app provider understanding and documenting the matching process, not just its outcome. Therefore, the matching algorithm must be transparent and demonstrate how the matching was conducted, allowing the app provider to assess the validity of consent. However, recording the privacy preferences of the data subject may infringe their privacy as with knowledge of these preferences, the controller can submit targeted requests aligned with those preferences, even though they intend to use it for different purposes.

The liability in case of personal data misuse takes on a distinct framework if the Pod provider is regarded as a separate data controller. In this scenario, the data subject gives consent to the Pod provider (Controller 1) for the storage and provision of personal data to third parties, under specific conditions. Following the data subject's instructions, the Pod provider subsequently makes the data available to the app provider (Controller 2). Each data controller must then ensure that the processing activities (storage, transfer, and further utilisation) are grounded in a valid legal basis [European Data Protection Board, 2020a]. Thus, the responsibility for errors in the matching process and for invalid consent could be jointly held by both controllers as per their mutual agreement. While the GDPR does not stipulate a specific legal form for such arrangements, the EDPB recommends documenting this agreement in a binding document, such as a contract accessible to the data subject [European Data Protection Board, 2020a].

5.4 Special categories of data and research exceptions

As discussed in preceding Sections, the stringent requirements for obtaining consent under the GDPR place significant burdens on data subjects. Requiring separate agreements for each app provider and for each specific purpose leads to repetitive consent requests. In the biomedical field, individuals may have less involvement in decision-making regarding their data compared to other sectors – the benefits from participation in biomedical research are often not immediate and may not directly impact the individual's personal circumstances. Consequently, individuals may be less inclined to make the effort of checking their Pod for new requests, reading information notices, and accepting/rejecting access requests, compared to sectors like information society services which include social media and streaming services. As such, in the context of research, various provisions indicate a more flexible approach to consent requirements or even suggest moving

away entirely from reliance on consent.

Recital 33 [2016b] states that broad consent is permissible for research purposes under specific conditions – *"data subjects should be allowed to give their consent to certain areas of scientific research when in keeping with recognised ethical standards for scientific research"* and while having the *"opportunity to give their consent only to certain areas of research or parts of research projects"*. However, the terms 'areas of research' or 'parts of research projects' are domain-specific concepts that are not further defined in the GDPR. The Global Alliance for Genomics and Health[8] develops distinct components and processes for health data sharing, including the Data Use Ontology (DUO)[9], a vocabulary that can be used to describe data use conditions and limitations for research data generated in the health, clinical and biomedical domain [Lawson et al., 2021, Rehm et al., 2021]. While such vocabulary contains concepts of health-related research purposes, links to other ontologies with disease taxonomies, and incorporates concepts for modeling projects and obligations related to data usage, e.g., need for ethical approval, collaboration with the study's investigator, or the obligation to return the study's results, it does not take into consideration data protection-related requirements, e.g., legal grounds for processing.

However, this provision outlined in Recital 33 [2016b] is not legally binding, is not mirrored in the actual text of the GDPR and it was strictly interpreted by the EDPS in its opinion on data protection and scientific research [European Data Protection Supervisor, 2020]. Furthermore, while the EDPS asserts that Recital 33 does not supersede the provisions mandating specific consent, it also suggests an assessment based on the data subject's rights, the sensitivity of the data, the nature and objective of the research, and relevant ethical safeguards. Concurrently, the EDPS also notes that if purposes cannot be precisely specified, data controllers could compensate by enhancing transparency and implementing safeguards. Outside of the EU, the UK government advocated for an influential role for broad consent in medical research within its proposal to amend the UK's Data Protection Act [UK Government, 2022]. This proposal was generally well-received, though some concerns were expressed regarding its potential for ambiguity and possible misuse.

Furthermore, the European Commission has proposed a regulation instrument for the health data domain, the European Health Data Space [2022f], which aims to depart from relying on consent for the secondary use of personal data in biomedical research. According to this proposal, a 'data holder', or a *"any natural or legal person, which is an entity or a body in the health or care sector, or performing research in relation to these sectors [...] who has the right or obligation [...] to make available, including to register, provide, restrict access or exchange certain data"* [2022f], is mandated to disclose both personal and non-personal data under specific conditions and for a limited set of purposes, including scientific research (Article 34.1(e) [2022f]), without requiring the consent of the data subject. Additionally, Article 33.5 of the EHDS proposal [2022f] appears to override national laws mandating consent by stipulating that *"where the consent of the natural person is required by national law, health data access bodies shall rely on the obligations laid down in this Chapter to provide access to electronic health data"*. The final version of this proposal, including the role of consent and its scope (broad or specific), is yet to be determined. However, this proposal has drawn criticism from both the EDPB and EDPS in a joint opinion document, which calls for further clarification on how national laws requiring consent will interact with the proposed European

[8] https://www.ga4gh.org/ (accessed on 9 March 2024)
[9] http://purl.obolibrary.org/obo/duo (accessed on 9 March 2024)

legislation [2022d].

Biomedical research presents challenges within the GDPR due to its unique combination of a stringent regulatory framework, as it involves processing health data, which falls under the GDPR's special categories of data, alongside a set of exemptions designed to facilitate research due to its societal significance.

5.4.1 A stricter regime for health data processing

GDPR's Article 9.1 [2016b] prohibits the processing of special categories of data, including health data. However, there are ten exceptions to this rule, one of which is explicit consent from the data subject. Nevertheless, the term 'explicit' lacks clarity, as it's unclear what distinguishes it from 'regular' consent, which requires a clear affirmative action or statement by the data subject. Further clarification is needed in the GDPR regarding the additional steps a controller should take to obtain explicit consent from a data subject [European Data Protection Board, 2020a]. The EDPB offers various examples of how explicit consent can be expressed. These include providing a written statement, or in the digital context, actions such as filling out an electronic form, sending an email, uploading a scanned document bearing the data subject's signature, or using an electronic signature. Another method mentioned is two-stage verification, where the data subject may receive an email from the controller requesting consent to process specific medical data. Upon agreement, the data subject is asked to respond via email with the phrase 'I agree', followed by receiving a verification link or an SMS message for confirmation [European Data Protection Board, 2020a].

Within decentralised frameworks such as Solid, various approaches can be employed for the purpose of expressing explicit consent. Depending on the Solid server chosen by users to host their Pod, an inbox container, akin to email inboxes found in other systems, may be automatically created when the user sets up the Pod. This container can serve as a platform to receive such requests as it is equipped with a specialised access control authorisation, allowing only the data subject to read its contents while permitting other users to write to it. However, due to the lack of standardisation across the Solid ecosystem, the presence of this container cannot always be guaranteed, or it may be named differently, leading to interoperability issues. A more sophisticated solution involves adopting a graph-centric interpretation of a Pod, wherein each Solid Pod functions as a hybrid, contextualised knowledge graph [Dedecker et al., 2022]. In this context, 'hybrid' denotes support for both documents and RDF statements, while 'contextualised' signifies the ability to associate each document and statement with metadata such as policies or provenance data. By accurately recording metadata, including context and provenance, multiple interfaces of the Pod can be generated as needed by various applications chosen by the data subject. In this scenario, requests can be seamlessly integrated into the graph without requiring hardcoded specifications in the application for where the requests should be written. These requests can then be visualised by the data subject using a Solid application or service compatible with this graph-centric approach. Additionally, the research conducted by Braun and Käfer [2022b] can be utilised to sign and validate resources carrying the 'I agree' statement of the data subject.

In summary, expressing explicit consent through pre-set polices poses challenges. Matching user policies, predefined in advance, with data requests is unlikely to meet the explicit nature of consent. Although matching can enhance transparency and assist individuals in decision-making, a separate

action of explicitly approving the use of personal data is required to meet the explicit requirement of consent.

5.4.2 A series of derogations for research purposes

In this Section, three distinct aspects, relevant to the domain of health research, are discussed: (i) the compatibility between data collection purposes and secondary reuse for research, (ii) exceptions from the right to information, and (iii) alternative exceptions, aside from consent, for processing special categories of data.

Secondary use for research As previously discussed in Section 5.3.2, related to the 'purpose limitation' principle and the assessment of compatibility, the GDPR states that *"data shall be collected for specific, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes"* (Article 5.1(b) [2016b]). As such, there is an assumption of compatibility between the purpose of collection and subsequent reuse, provided that the personal data processing for scientific research purposes appropriately implements safeguards to protect the rights and freedoms of the data subject (as outlined in Article 89.1 [2016b]). It is crucial to highlight that the prohibition against processing personal data for incompatible purposes differs from the requirement of purpose specificity, and an exception does not alleviate the need for a specific purpose. Moreover, regardless of compatibility, the data controller must rely on consent or another legal ground to process personal (health) data. However, there is one provision in the GDPR preamble that questions the distinction between these two requirements – *"The processing of personal data for purposes other than those for which the personal data were initially collected should be allowed only where the processing is compatible with the purposes for which the personal data were initially collected. In such a case, no legal basis separate from that which allows the collection of the personal data is required"* (Recital 50 [2016b]). This appears to challenge the separation between the 'purpose limitation' and the 'lawfulness' principles. This intersection, and its implications for decentralised data-sharing ecosystems such as Solid, needs to be further investigated.

Exceptions to the information obligations In Section 5.3.1, particularly in the "Identifying the data controller" paragraph, the information obligations outlined in Articles 13 and 14 of the GDPR [2016b] are explored, with a focus on the timing of when information must be provided to the data subject. In particular, Article 14 provides an exception for cases where personal data are processed for research purposes and have not been obtained directly from the data subject. This exception may be relevant to Solid, considering that not all personal data stored in Solid Pods originates directly from the data subject – it may be generated by app providers, Pod providers, other users, or agents. Furthermore, according to Article 14.5, if (i) providing information is impossible or would require disproportionate effort, or if doing so is likely to render impossible or seriously impair the achievement of the processing objectives, and (ii) the conditions and safeguards specified in Article 89 [2016b] are met, the information requirements outlined in Article 14 are inapplicable. The compliance of Solid-based data exchanges with these conditions and safeguards in place will need to be evaluated on a case-by-case basis, depending on the context and the data access request. However, it is probable that these conditions will be fulfilled only in exceptional cases rather than as a standard practice, and if they are met, the data controller *"shall take appropriate measures to protect the data subject's rights and freedoms and legitimate interests, including making*

the information publicly available". As such, further research is needed to explore the role of Solid's notification system, as well as other mechanisms, to act as appropriate measures to safeguard the rights of the data subject.

Alternative legal bases beyond consent In addition to explicit consent, GDPR's Article 9.2 [2016b] outlines other exceptions to the prohibition on processing special categories of data. Article 9.2(j) is particularly pertinent to this discussion because it pertains to the processing of personal data for health research. This point permits the processing of health-related data when it is necessary for scientific research in accordance with Article 89.1 [2016b], as long as it is based on European or national law. Such processing must be proportionate to the intended purpose, uphold the essence of the right to data protection, and include appropriate and specific measures to safeguard the fundamental rights and interests of the data subject. Consequently, the applicability of this exception hinges on the identification of a European or national law that can justify the processing of personal data. If the processing falls within the scope of such legislation, explicit consent from the data subject is not required.

As such, from this Section is possible to conclude that the exemptions for processing personal data for scientific research hinge on the adoption and use of suitable safeguards. According to GDPR's Article 89.1, these safeguards center around upholding the 'data minimisation' principle and include practices like pseudonymisation and methods that prevent the identification of data subjects. Subsequent research could explore whether PIMS, such as the Solid with an OAC-based matching system, could serve as a safeguard in this context as by nature data minimisation can be provided using OAC's fine-grained access policies.

5.5 Ethical challenges of controlling data and reclaiming control over it

Advancements in data-driven innovations are poised to drive further economic and societal progress [Jacobides et al., 2019]. The analysis, sharing, and reuse of data have led to transformative changes in business models and government processes, enabling them to capitalise on these practices. As discussed in the previous Sections, these changes propelled policy initiatives implemented by various governments globally. In particular, the EU is actively engaged in this transformation, exemplified by the European Commission [2020] commitment[10] to shaping "A Europe fit for the Digital Age". Whether it is a prominent Big Tech firm headquartered in the United States, a major data intermediary in the EU, or a state-controlled entity in China, contemporary data practices face scrutiny from diverse sectors of society, spanning individuals, non-governmental organisations (NGOs), academics, and governmental bodies. Such distrust in digital services has been called into question [Waldman, 2021], prompting individuals to ponder who should they trust their data with.

Amidst this trust crisis, technology has emerged as a potential solution, in particular self-sovereign PIMS [Chomczyk Penedo, 2021], as discussed in Section 5.1. These models empower users to

[10]The European Commission's strategy and related documents are available at https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age_en (accessed on 10 March 2024)

directly control their data, dictating the terms of access and usage, and have been gaining the support of policymakers, in particular in Europe, with the European Commission supporting the creation of common European data spaces [2022a]. Moreover, it could be argued that the EU is strategically investing in these technologies to foster more democratic and participatory data practices, and enhance confidence in data-intensive operations by advocating for technologically robust systems that reduce reliance on the reputation of individual firms, thus mitigating power imbalances between data subjects and controllers [European Commission, 2020].

The literature exploring the concept of trust is extensive, yet complex due to varying interpretations. De Filippi et al. [2020] distinguished trust from confidence, noting that trust is rooted in personal vulnerability and risk-taking, while confidence is based on internalised expectations stemming from knowledge or past experiences. As such, in this Section, the interest of data subjects in technologies that provide insights into how their information is integrated into real personal data handling processes is studied as a vehicle of trust, given their general apprehension regarding the processing actions of data controllers over personal data. As visible in the previous Sections, the personal data regulatory framework in the EU is designed to address imbalances or vulnerabilities between multiple parties by revealing potential risks and resulting harms, aiming to leverage consent as a catalyst for the data-driven economy [Chomczyk Penedo, 2022]. Simultaneously, they aim to furnish essential information to individuals making decisions, facilitating informed choices Ben-Shahar and Schneider [2014]. Furthermore, from an ethical standpoint, several norms emerge that should guide the conduct of individuals with whom information is shared to ensure trustworthiness. These norms encompass sincerity, competence, and the appropriateness of the entrusted task [Hawley, 2019].

Considering the myriad of factors influencing both trust and confidence, the analysis in this Section focuses on (i) transparency as a crucial prerequisite for the functioning of decentralised PIMS, (ii) the relevant EU regulatory framework on personal data, and (iii) an ethical debate concerning data control, as outlined in Bodó's framework for mediated technological trust. The emphasis on transparency stems from three primary considerations:

- from a regulatory viewpoint, transparency stands as a fundamental principle within personal data protection regimes, often integrated alongside lawfulness and fairness, as exemplified in GDPR's Article 5.1(a);
- transparency encompasses both its *ex-ante* and *ex-post* components, with the latter including the issue of explainability [Felzmann et al., 2019];
- transparency offers the potential to demystify the 'black box' nature of many AI systems, enabling the identification of potential biases towards vulnerable populations [Pasquale, 2015].

As illustrated by case law from supervisory authorities, the intricate nature of data processing activities has proven challenging for data controllers to articulate in straightforward terms, especially when relying on limited attention resources from data subjects [European Data Protection Board, 2020a]. The dearth of actionable information, to understand data handling practices, poses a risk to fostering trust among involved parties. As a result, individuals are endeavoring to reassert control over their data and restrict its usage by such entities, also by looking at new data governance schemes such as PIMS or other data intermediaries [Craglia et al., 2021, Papagiannakopoulou et al.,

2014].

As such, the concept of 'control' gains particular importance as users require someone to trust in order to reclaim control over their data in the digital era. Emerging data governance models are coupled with legal frameworks to assist data subjects in asserting their agency. For instance, in the data cooperative model (which is regulated by the DGA), cooperatives act as trustees overseeing data on behalf of data subjects, thus enabling data subjects to maintain democratic control over their data. In such governance frameworks, establishing a relationship of trust between cooperatives managing data and data subjects is paramount. In certain instances, trustees may need to consult with data subjects, providing agreements and contracts to inform them. Meanwhile, data subjects can articulate their preferences and determine how to share their data and for what purposes [Craglia et al., 2021].

Data cooperatives and other intermediaries (will) play a pivotal role in empowering data subjects to maintain control over their data and reassert their ethical standing in the digital era. Specifically, personal data sovereignty offers a significant return to more democratic and egalitarian governance, allowing individuals to reclaim control over their personal data [Craglia et al., 2021, Giannopoulou, 2023]. In theory, these systems should restore personal autonomy and uphold classical liberal values by fostering trust-based relationships. Furthermore, drawing from our current democratic experiences can offer valuable lessons to avoid repeating the same mistakes made in the past two centuries. During this time, a substantial portion of the population, particularly in the Global South, suffered from neglected rights due to inadequate governance safeguards. For instance, democratic failures in Latin America over the last 50 years, stemming from regime changes, economic crises, or environmental catastrophes, have led to the absence of robust governance mechanisms to address such challenges. One illustrative example is the impact of the last Argentinian military dictatorship, which significantly altered the identities of numerous individuals who were abducted as children and placed with new families, effectively erasing their true identities. In response, collective organisations emerged to address this injustice, recognising the vulnerable position these individuals were placed in and their limited ability to resist and reclaim their true identities [Gesteira, 2014].

Despite the critical role of trust in upholding the autonomy and agency of data subjects [Ben-Shahar and Schneider, 2014], the methods currently employed to foster trust remain contentious, and unresolved societal issues persist in digital services and emerging digital intermediaries [Carovano and Finck, 2023]. Given the practical nature of the issues at hand, including how to practically approach trust, establish trust relationships between data subjects and data intermediaries, and identify the necessary conditions for fostering trust, a public Think-In event was organised in the context of the PROTECT project. In these events, individuals were convened to explore the implications of governing personal data spaces through decentralised PIMS or trusted data intermediaries. With the "citizens' Think-In" approach, there is a public discussion focused on the opinion of individuals, which encourages direct participation from attendees. In particular, through small-scale group discussions, a Think-In offers a platform for individuals from diverse backgrounds to deliberate and exchange views on current societal issues stemming from advancements in Science, Technology, Engineering, and Mathematics (STEM) fields[11].

[11]Information regarding the organised PROTECT Think-Ins and respective results is available at https://w3id.org/people/besteves/phd/thinkin (accessed on 11 March 2024)

While the comprehensive outcomes of the Think-In process will not be included in this Thesis as a contribution, it is worth noting that the general public exhibited sensitivity toward the ethical considerations regarding whom to trust and the significance of transparency in such contexts. Citizens emphasised the importance of preventing the GDPR from turning into a mere 'tick-box' compliance exercise, similar to the current format of privacy notices which result from deploying template privacy notices for distinct data processing activities. Furthermore, there was a call for increased disclosure and oversight concerning the practical and beneficial utilisation of personal data, highlighting the importance of meaningful transparency in fostering trust among parties involved in such sensitive data exchanges.

To conclude, the insights derived from the Citizens' Think-In discussion offer a valuable foundation for considering the integration of transparency into data access agreement terms for personal data vaults, presented in both machine-readable and human-readable formats. As such, the proposed vocabulary work, described in Chapter 4, represents a first step to offer said transparency for data subjects, containing both machine-readable and human-readable descriptions of concepts. This approach enables data subjects to better comprehend and manage the expression of policy terms, and empowers data controllers and data subjects to navigate the intricate data-sharing landscape of the platform economy with greater control vested in the data subject.

Part III

ALGORITHMS & USE CASES

Chapter 6

Design of a Policy-based Algorithm for Access to Decentralised Personal Datastores

The content of this Chapter has already been partially included in the articles published during this Thesis [Esteves et al., 2021, 2022e, Pandit and Esteves, 2024, Esteves et al., 2022d].

The source code produced during the development of this chapter is stored at:

- https://w3id.org/people/besteves/sope/repo
- https://w3id.org/people/besteves/access-right/api
- https://w3id.org/people/besteves/access-right/solid
- https://w3id.org/duodrl/repo

This Chapter describes an architecture for a legally-aligned, decentralised personal data stores ecosystem, including the description of a policy matching algorithm and data access agreement generator prototype that uses and extends the developed vocabularies to deal with the specific requirements of health data sharing. This Chapter builds upon the vocabularies described in Chapter 4 to bring decentralised datastore environments, such as Solid, closer to being compliant with data protection law in Europe, by improving its transparency and accountability mechanisms through interoperable, machine-readable information which can be recorded and consulted by data subjects, data controllers and other interested parties. The prefixes and namespaces used in the Listings in this Chapter are explicitly defined in the Namespaces list.

Section 6.1 describes a detailed decomposition of the architectural building blocks for a legally-aligned, decentralised personal datastores ecosystem using the C4 graphical notation model [Brown, 2015]. System context, container, and component architectural diagrams are provided to report the proposed system, and a sequence diagram is provided to demonstrate how they work together.

Section 6.2 provides a description of the design of algorithms for instantiating user requirements and preferences as user offers and for generating data access agreements, the outcome of matching

data requests with user offer policies. Furthermore, the Solid ODRL access control Policies Editor (SOPE), an RDF and ODRL-agnostic policy editor to define and store OAC policies in Solid Pods, is described, as well as a REST API service that uses the results of the policy matching algorithm to facilitate the exercising of data subject rights, in particular of the right of access.

Section 6.3 discusses the implementation of a proof of concept implementation of the DUO-based algorithms for the specific use case of health datasets shared for research purposes.

To conclude, Section 6.4 evaluates the developed algorithms and PoC by assessing ontology quality, including pitfall detection and alignment with FAIR principles, and by comparing the proposed policy-based algorithm with DUO's and Solid's existing access control systems. Concluding remarks concerning the architecture, algorithms, and PoC implementation are also presented.

6.1 Architecture for the deployment of a policy matching algorithm for access control

In this Section, a detailed decomposition of the architectural building blocks for a legally-aligned, decentralised personal datastores ecosystem is modelled and documented using the C4 graphical notation model [Brown, 2015]. The C4 model is a formalisation used to visualise software architecture, based on the 4+1 View Model of software architecture [Kruchten, 1995], which has evolved over the years to showcase different views of software components, each of which addresses a specific set of issues, inspired by the Unified Modelling Language (UML). The main objectives of this model are (i) to simplify the description and understandability of software systems for software developers and (ii) to decrease the gap between source code and software architecture modelling. The four visualisations of the C4 model have the subsequent goals:

- The *System Context* diagram serves as an initial framework for illustrating and documenting a software system, providing an overview that allows for a comprehensive understanding of the system's environment. It typically features the system to be decomposed as a central entity, surrounded by its users and other interconnected systems. The emphasis lies in presenting a broad perspective of the system landscape for non-technical audiences, with less emphasis on intricate details. The primary focus is on identifying people, e.g. users or roles, and software systems, rather than delving into technical specificities such as technologies or protocols.
- The *Container* diagram offers an overview of the software architecture's structure at a high level, delineating the allocation of responsibilities within it. Furthermore, it illustrates the primary technological selections and elucidates the communication channels between components, e.g., server-side web applications, mobile apps, or file systems.
- The *Component* diagram illustrates the decomposition of a container into various components, elucidating the purpose of each component, and the technological or implementation details involved.
- The *Code* diagram is an optional visualisation, recommended only for critical components, that zooms in on each component to illustrate its implementation as code, employing UML class diagrams, entity relationship diagrams, or comparable methods.

In the following Sections, the architectural model for a legally-aligned, decentralised personal datastore will be discussed in detail through context, container, and component diagrams. A special focus will be given to the proposed personal datastore server implementation, and the agreement generator and datastore containers and their components.

6.1.1 System context modelling of a decentralised data sharing ecosystem

The *System Context* diagram in Figure 6.1 illustrates how decentralised personal datastores interact with legal entities and/or natural people and with internal and external software systems at a very high level. This diagram depicts the decentralised personal datastore server at the centre with no details of its containers, surrounded by all its interacting systems and actors. The depicted server architecture is aligned with the existing Solid protocol specification [Capadisli et al., 2022] and existing Community Solid Server and Enterprise Solid Server implementations.

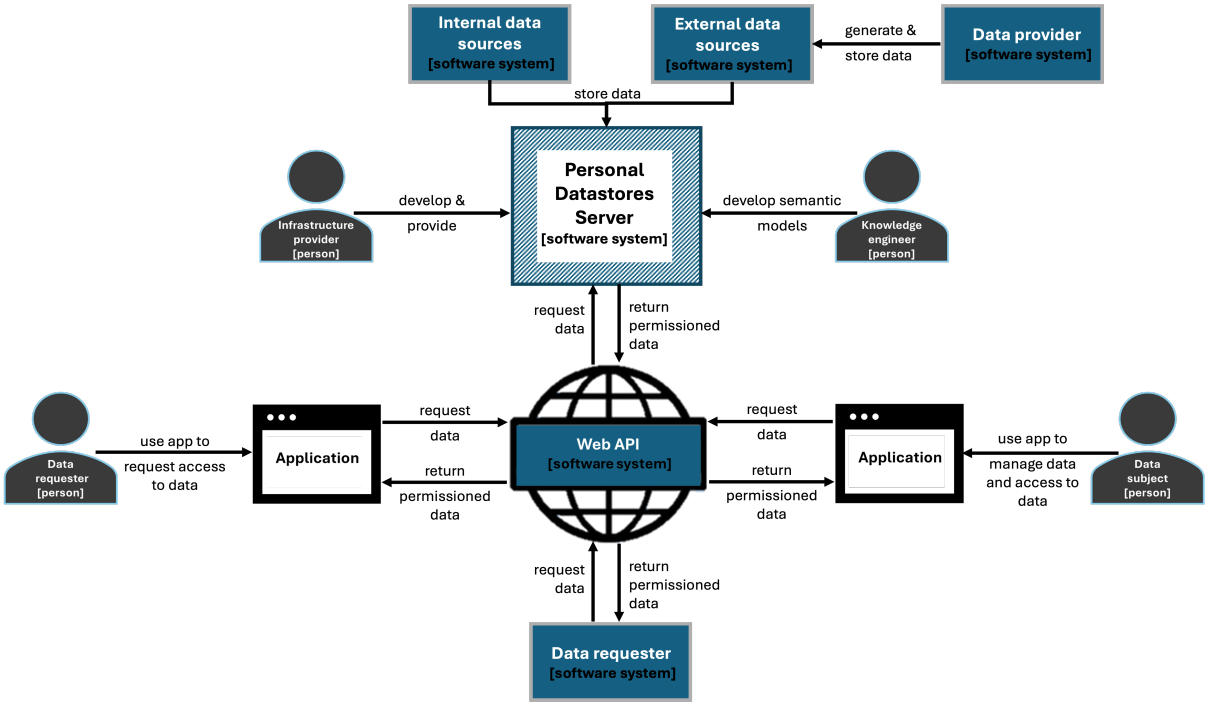


Figure 6.1: System context diagram of a decentralised data sharing ecosystem.

In this diagram, the *Infrastructure provider* and the *Knowledge engineer* develop and provide services and semantic models for the main software system described in this Section, the *Personal Datastores Server*. Datastores hosted in this server are populated by internal and external data sources, such as temporal and spatial data sources, respectively. The server also provides the access control layer of the ecosystem, allowing data requests and the return of permissioned data to be consumed by Web APIs. Such APIs can be used by *Data requesters* to consume and generate data, and also to feed Web applications that are used by individuals to request and use data. Moreover, data subjects can also use applications to manage their data and who/what can have access to it under which conditions.

The *Personal Datastores Server* depicted in this Figure will be further decomposed into containers

in the next Section.

6.1.2 Container modelling of a decentralised personal datastore server

The *Container* diagram in Figure 6.2 illustrates current and proposed containers within a decentralised personal datastores server. In a completely decentralised scenario, each container could be developed and/or provided by a distinct *Infrastructure provider*. The *Notifications*, *Authentication* and *Authorisation* containers are already part of the Solid protocol specification [Capadisli et al., 2022] and are not further updated in this Thesis. To ensure compatibility with the current access control mechanisms, the *Offer Instantiation* and *Agreement Generator* containers are treated separately from the *Authorisation* container. In this way, systems that currently rely on WAC or ACP to provide access to data can continue to work efficiently, while use cases that involve personal data can use OAC policies to have data access agreements aligned with European data protection law. ACL and ACP authorisations can be generated, when an agreement for data access has already been reached, for automated access to data when a legally-aligned agreement for access is already in place. The *Notifications* container can also be updated to support the new access control system proposed in this Thesis, however, this is out of the scope of this contribution.

Furthermore, while an intricate part of the Solid protocol already, the *Datastore* container is further expanded in this Thesis with additional components related to policies and provenance metadata. The vocabularies developed in Chapter 4 should be further developed and maintained by a *Knowledge engineer* and are used to specify data, provenance, and policies, as well as to feed the *Offer Instantiation* and *Agreement Generator* containers. The *Offer Instantiation* container gathers and instantiates ODRL offers, such as the one in Listing 4.2, which are fed to the *Agreement Generator* container. This instantiation process takes the user's preferences and requirements to create a context-appropriate offer policy, i.e., directly applicable to the data request. While the implementation of an offer instantiation algorithm for specific data requests is not conducted in this Thesis, in Section 6.3, there is a description and implementation of the construction of odrl:Offer instances for a particular use case related to health data-sharing. The *Agreement Generator* depicted in Figure 6.2 will be further decomposed into components in the next Section.

6.1.3 Component modelling of a datastore and an agreement generator

The *Component* diagram in Figure 6.3 further decomposes the *Datastore* and *Agreement Generator* containers within a decentralised personal datastores server. As previously discussed in Sections 4.2 and 4.3, for a legally-aligned decentralised ecosystem where personal data is exchanged according to the requirements of the GDPR, information regarding the entities (e.g., providers, developers), infrastructure, legal roles, policies, notices, registries, and logs is necessary to understand and establish responsibilities and accountability within the whole ecosystem. As such, beyond keeping *Data* in distinct formats, *Datastores* are expected to keep (i) *Policies*, including user requirements and preferences, copies of data requests and records of data access agreements, (ii) *Notices*, e.g., to declare the providers/developers of apps, service, or other infrastructures, as well as to describe their data processing practices, (iii) *Logs*, related to actions on data, policies, or security issues, which can be used by external auditors to perform their duties, (iv) *Registries*, for collective and convenient access to data, and (iv) other *Provenance* metadata, all needed to have a trustful and

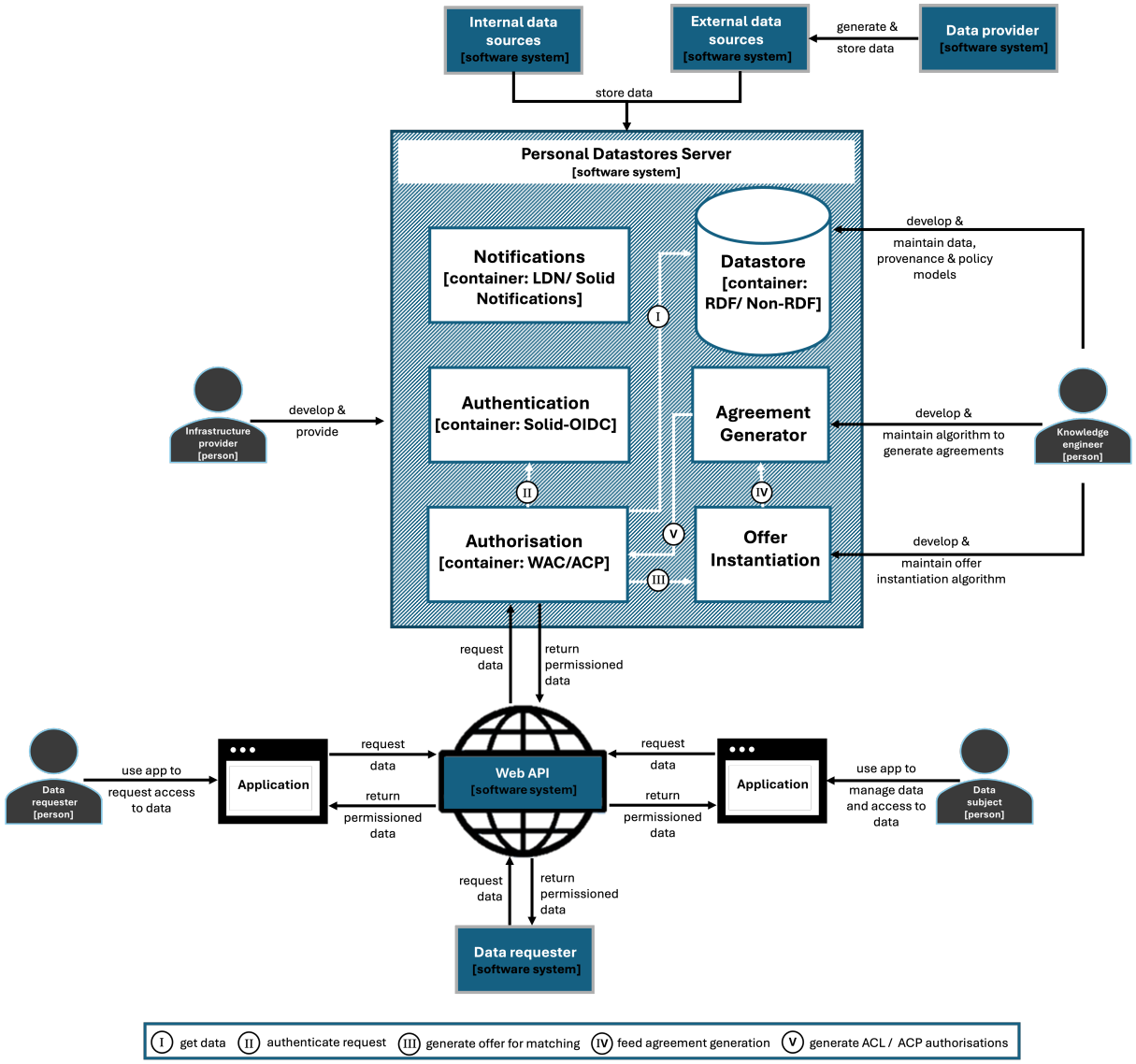


Figure 6.2: Container diagram of a decentralised personal datastore server.

sustainable decentralised data sharing ecosystem. As such, the semantic models produced in Chapter 4 play an important role in having interoperable datastores where data can be consumed by multiple parties in a lawful manner, even if said parties resort to the use of different providers.

When it comes to the architecture of the *Agreement Generator*, the following components were identified:

- The *Request Validator* component has the main goal of validating RDF-based data requests against the conformance conditions established in Section 4.3.3, using technologies such as SHACL [Knublauch and Kontokostas, 2017] or ShEx [Prud'hommeaux et al., 2019]. This component is of fundamental importance to ensure that all legally required data is present in the data request and to guarantee that the data request can be fed to the policy matcher in the expected format.

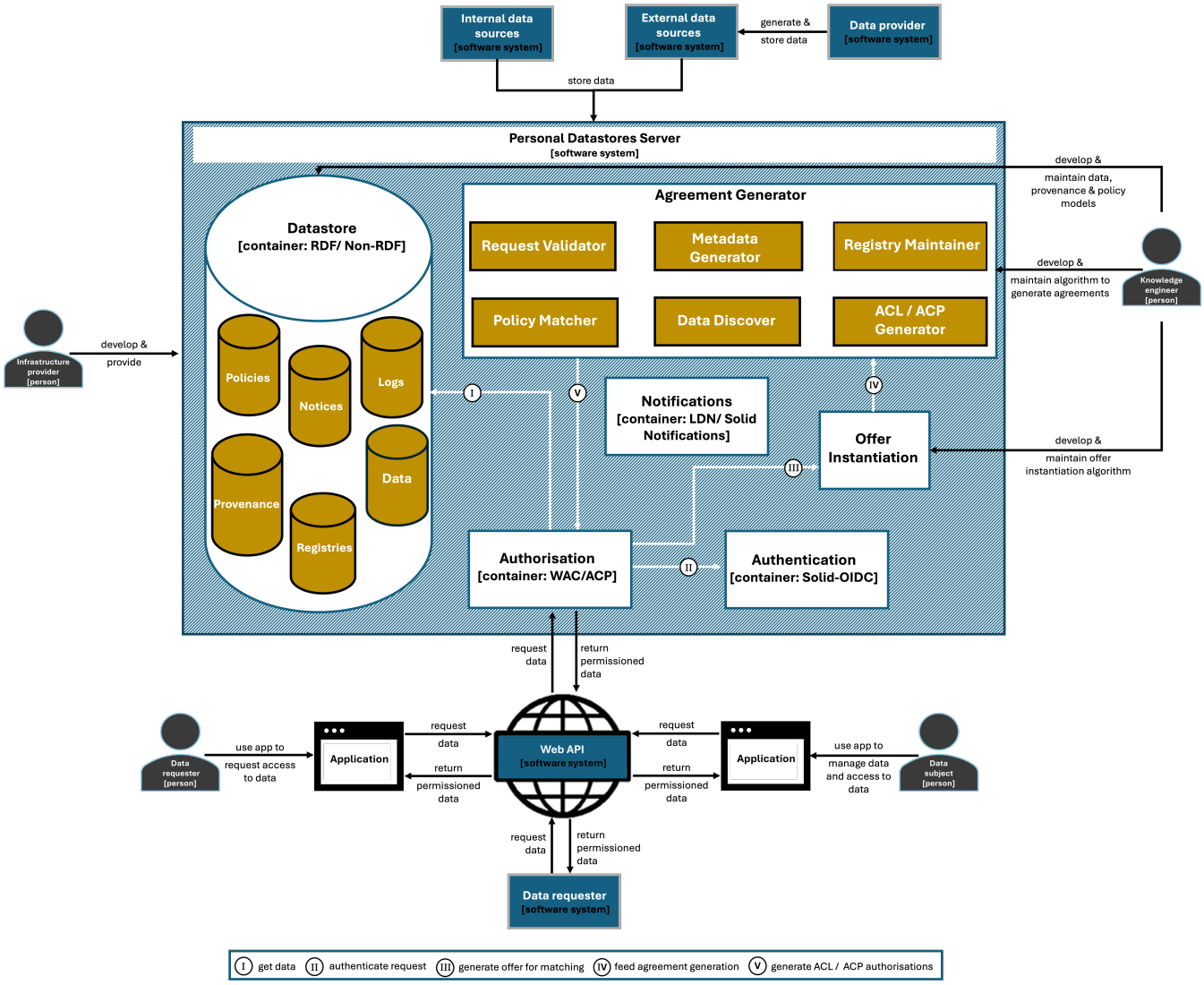


Figure 6.3: Component diagram of a datastore and an agreement generator.

- The *Metadata Generator* and *Registry Maintainer* components have the main goal of generating and storing logs and other metadata related to entities, apps, services, or Pods and maintaining updated registries of access control authorisations, availability of data categories, supported schemas for data, and relevant policies, apps, services, users, or agents.
- The *Data Discover* component's main task is to provide the server with data discovery capabilities in order to more easily find where certain types of data, policies, or provenance metadata are stored, including storage locations outside the server.
- The *Policy Matcher* component, further specified in Section 6.2, has the main goal of matching the existing user offers with incoming data requests and generate data access agreements. These agreements are then fed to the *ACL/ACP Generator* component for seamless integration into the current Solid ecosystem.

Indeed, to understand the detailed functionality of the proposed personal datastores server, Fig-

ure 6.4 presents a sequence diagram to demonstrate how its containers, and more specifically its components, interact with each other when a data requester uses an app to petition a certain type of data to be used for a particular purpose.

When data requesters want to have access to a certain resource or data type, they can either use an app to do it, as in Figure 6.4, or directly do a request through the Web API. When using an app, the request is then transmitted to the Web API and in turn, the requester must be authenticated to get access to non-public data. After the authentication process goes through, or also in the cases where requesters cannot successfully log in, an identity log is recorded. Additionally, information about the data requester is also recorded in the user registry of the data subject. If a previously-given authorisation is still valid, then access to data can be given, however, in Figure 6.4, this authorisation does not exist, a fact which could also be kept in an access control log.

Therefore, as there are no ACL nor ACP authorisations in place, an offer with the data subject's policies should be instantiated in order to feed the agreement generator. This instantiation is followed by the validation of the data request, to ensure that it is in the correct format, and both are fed to the policy matcher component for the generation of a data access agreement. A log of the policy matching and its result should be kept and the resulting agreement should be added to the policy registry. In case the result is negative, access to data is prohibited, contrarily to the scenario depicted in Figure 6.4. On the other hand, if the result is positive, access to data is permitted, under certain conditions, and an ACL and/or ACP authorisation is generated. Afterwards, an access control log is recorded and the new authorisation is kept in the access control log for future usage. Finally, the data discover mechanism is used to find the data being requested and this data is returned to the Web API and next to the app for the data requester to have access to it.

In the next Section, the algorithms for offer instantiation and for policy matching are further explored, with the goal of generating data access agreements that are aligned with legal requirements.

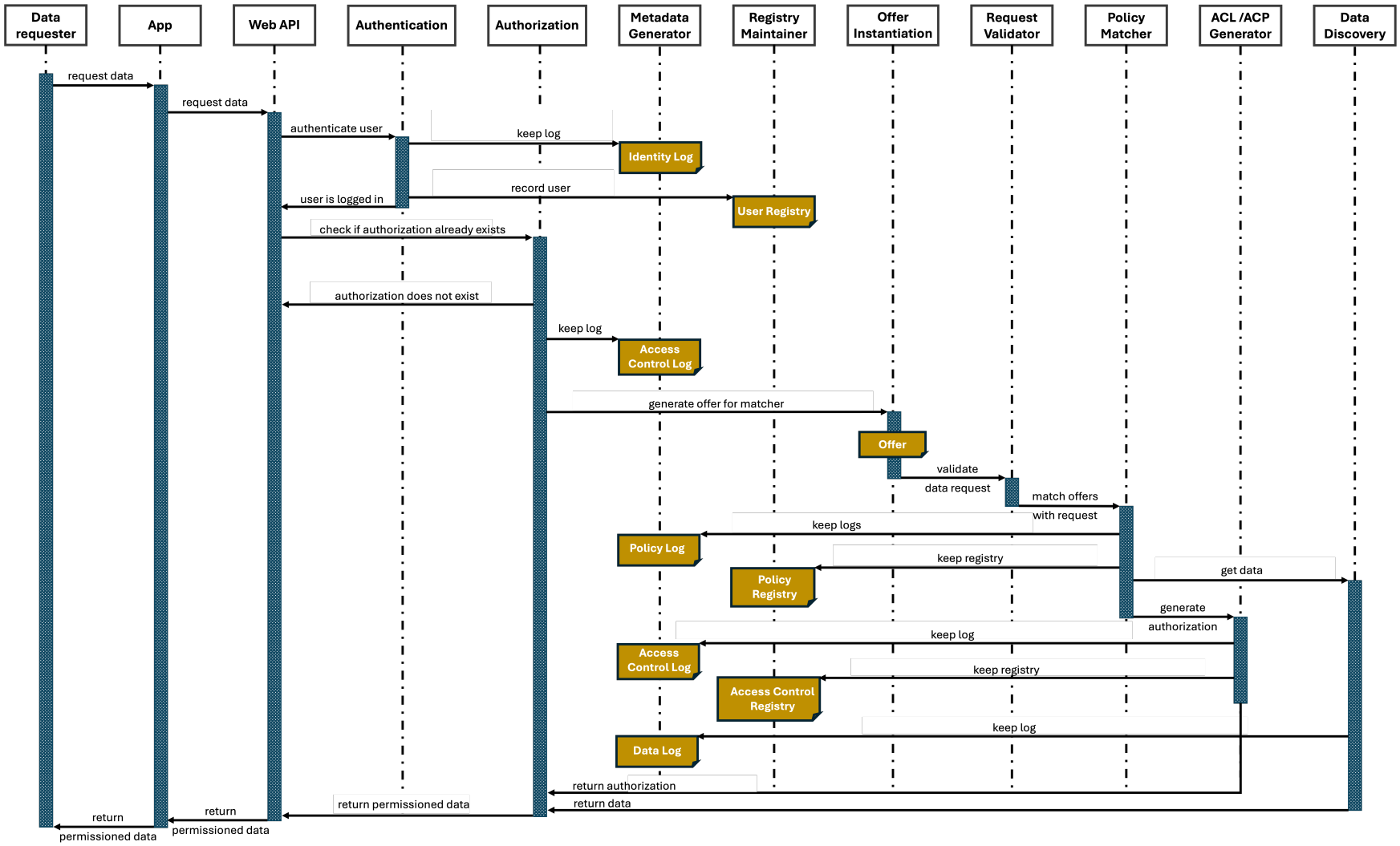


Figure 6.4: Sequence diagram of data access request using proposed architecture.

6.2 Design of a policy matching algorithm for generating data access agreements

In this Section, a detailed overview is given of the proposed algorithms for offer instantiation and policy matching towards the generation of a common data access agreement. Moreover, a policy editor to define and store OAC policies in Solid Pods is described, as well as a REST API service that uses the results of the policy matching algorithm to facilitate the exercising of data subject rights.

6.2.1 Development of an OAC policy editor

This Section features an ODRL editor designed to define and save RDF policies, to enable the granting of access to personal data stored in Solid Pods. RDF policies are articulated using the OAC specification developed in Section 4.2. As such, with such an OAC editor, data subjects can create intricate, detailed policies that adhere to GDPR stipulations concerning personal data processing without the burden of knowing ODRL or even RDF.

SOPE – the Solid ODRL access control Policies Editor SOPE is a Solid-based app for data subjects to define and store ODRL policies, based on the OAC specification, on Solid Pods. Detailed instructions on how to install, launch, and use the app are available on the source code repository[1]. To use it, data subjects must already be in the possession of a Solid Pod and a WebID to be able to log into their Pod and store policies there, using SOPE. Once logged in, users can select the type of policy they want to model, as well as choose the types of personal data and purposes to which the policy applies. Additionally, users can also select what type of access they which to provide for said data type and purpose. Finally, the prototype interface allows users to generate and store the ODRL policy's RDF in the Pod, without the need for users to be knowledgeable about OAC, ODRL, or DPV. SOPE also stores policy logs and updates the policy registry in the user's Pod, using the PLASMA vocabulary to model such information, however, in the future, such documentation should be generated by the server to ensure interoperability (by not depending on the app's own implementation, which could diverge from app to app). All generated information, e.g., policies, logs, and registry, is stored in a private location within the Pod – only authorised users, apps, or services will have access to it if desired by the data subject. Figure 6.5 presents a screenshot of SOPE's interface.

SOPE coverage, maintenance, and future work SOPE is published and archived according to the methodology described in Section 3.7.3. Furthermore, SOPE's source code is hosted at https://w3id.org/people/besteves/sope/repo, under the CC-BY-4.0 license. A live demonstration of the app's features is also available at https://w3id.org/people/besteves/demo/eswc22. The repository can also be used by SOPE users to suggest new features to be added to the app and to report bugs through GitHub Issues. Currently, SOPE's app coverage encloses terms from DPV's taxonomies of purposes, personal data categories, and processing operations. As future work, SOPE can be extended to include all terms present in the previously mentioned DPV's taxonomies, as well as to cover all constraints defined in the

[1] https://w3id.org/people/besteves/sope/repo

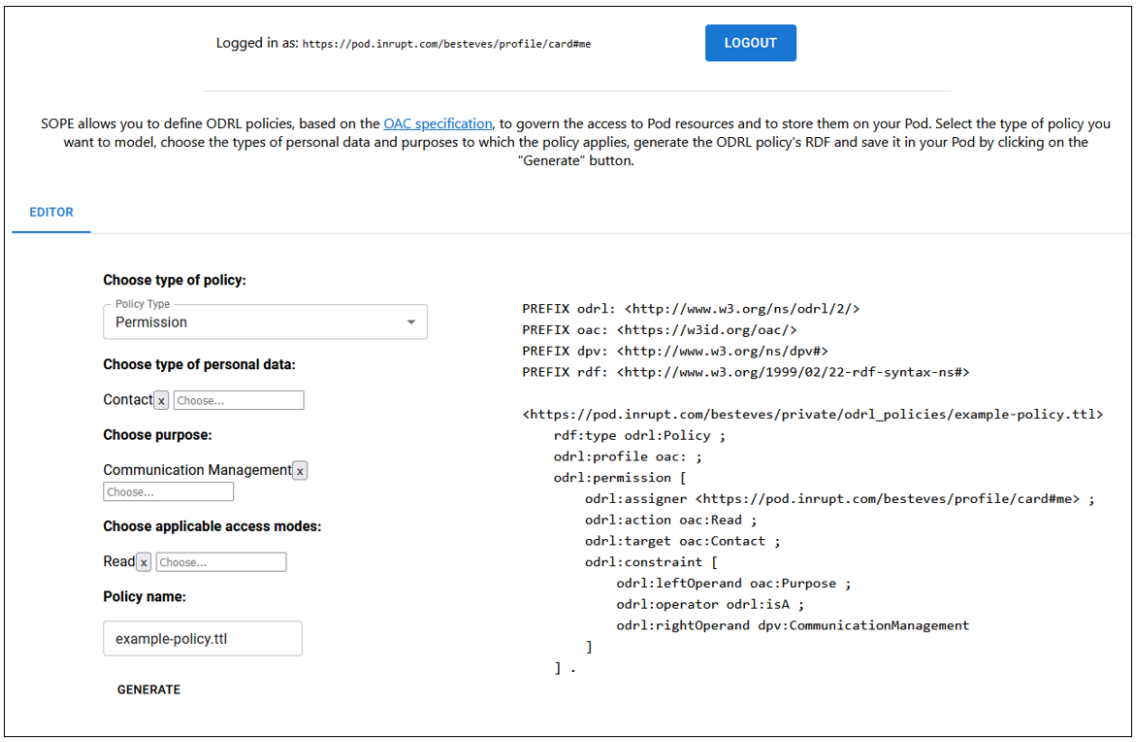


Figure 6.5: Screenshot of SOPE, a Solid app for editing OAC-based policies.

OAC profile, e.g., restrict legal bases and recipients, or specify the technical and organisational measures used by data controllers to ensure the secure processing of personal data. Moreover, with such an extension, SOPE could also be used by data controllers to detail their privacy policies. Additionally, user studies should be performed to assess the design choices included in the editor, as well as to understand what type of additional controls people want to have on top of what is legally mandated, e.g., temporal constraints or duties for the data controller to fulfil prior to data access.

6.2.2 Data subject policies as `odrl:Offer`s

Data subjects can express policies for specific resources, containers of resources, specific personal data types, or even for data they have not produced yet. As such, at the time of instantiation, the incoming data request should be used to filter the data subject's policies that apply to that particular request. This is to ensure that only the needed policies are shared and no more, aligned with the 'data minimisation' principle in GDPR's Article 5.1(c) [2016b], as user policies should also be considered personal data and be treated as such. Therefore, the resulting `odrl:Offer` instance contains the union of all pertinent user policies, which can be used to match against incoming data requests in order to generate data access agreements for certain resources or data. Furthermore, this policy can be used as metadata that accompanies the data being accessed/shared, so that data controllers can keep a copy of the conditions under which they can use the data. In essence, such a system improves trust and accountability in decentralised data-sharing ecosystems as the policy can travel with the data, with provenance information on who generated the data, who generated the policy, and who attached it to the data. Even though malicious agents can perform

prohibited actions, such as making copies of data when only read-access to data was allowed, with the documentation of access conditions being stored on the datastore of the data subjects, they can use these policies and metadata to file a complaint in the court of law for data misuse. Moreover, when formulating offers, each distinct rule merged into the policy is preserved as an individual rule to facilitate the matching with data requests. This preservation of individual rules also enables their individual annotation with provenance metadata, e.g., their origin or whether they are negotiable or non-negotiable rules.

Taking this into consideration, the instantiation of the `odrl:Offer` should follow the subsequent algorithm:

1. For a given personal datastore, retrieve all user preferences and requirements recorded in a datastore policy registry.
2. Filter out duplicated policies.
3. Filter out policies that do not match with any terms of the data request.
4. For each retrieved policy, fetch relevant `odrl:Permission` and `odrl:Prohibition` rules and merge them in a single `odrl:Offer`.
5. Permissions and/or prohibitions associated with an OAC requirement or an OAC preference should be associated with the term `dpv:Required`, in case said rule is a requirement, or the term `dpv:Optional`, in case it is a preference, using the `dpv:hasContext` property.
6. A link to each 'original' policy is maintained in the final `odrl:Offer` by using the `dcterms:source` property.
7. Add provenance information to the `odrl:Offer`, e.g. `dcterms:issued` for when the offer was instantiated and `dcterms:creator` for the issuer of the policy.

The previously described Listing 4.2 presents an example of a result of the offer instantiation algorithm, generated from two existing, relevant policies, as indicated by the `dcterms:source` property, based on an OAC requirement and OAC preference policies, as expressed by the `dpv:hasContext` property.

6.2.3 Policy matching outcomes as `odrl:Agreement`s

As mentioned in the previous Sections, the instantiated user policies are one of the few bits, representing the will of the data subject, that must be fed to the policy matching algorithm to generate the data access conditions to certain data or resources. In addition to those, data request policies, expressing the needs and purposes of the data requesters, as well as other contextual information, e.g., date or time of the day, must also be passed on to the algorithm in order to reach a data access agreement that benefits and satisfies both parties.

The recorded outcomes, resulting from the matching process where access to data needs to be either permitted or prohibited, are instances of `odrl:Agreement`. Within these agreements, specific ODRL terms play a crucial role in specifying who has granted or denied access (`odrl:assigner`), to whom (`odrl:assignee`), for what resources (`odrl:Asset`), and

the associated conditions for access (`odrl:Rule`). Moreover, the rules referenced within an agreement mirror the specific rules outlined for a dataset, i.e., through an `odrl:Offer`, and within a request, i.e., through an `odrl:Request`). As such, by being derived from these rules, an agreement should explicitly reference them to assist in the explainability of the algorithm, e.g., in case the data subject wants to know why a certain data access-related decision was made. An example representation of a data access agreement, depicted as an `odrl:Agreement` between two parties to read the Pod's data subject age data for academic research, is provided in Listing 4.4.

Taking this into consideration, the generation of the `odrl:Agreement` should follow the subsequent algorithm:

1. Retrieve the data requester's `odrl:Request` and the user policy's `odrl:Offer`.
2. Match the `odrl:Offer` with the `odrl:Request`.
3. Record the outcome of the matching algorithm, where the `odrl:target` property specifies the data to be accessed, and the `odrl:assignee` and `odrl:assigner` properties identify the requester and the data subject, respectively.
4. If the matching result is positive, i.e., the request and the offer are compatible, then access is permitted by employing a permission with constraints on the requested purpose for access, along with any additional constraints such as legal bases, identity providers or recipients. If access is denied, similar information is included in the policy as a prohibition.
5. Utilise the `dcterms:references` property to associate the agreement with the `odrl:Offer` and `odrl:Request` that were used to generate it.
6. Include provenance and other relevant information, such as the `dcterms:issued` property, to document the creation and acceptance of the agreement among the parties.

The matching process described in step 2 of agreement generation algorithm operates by comparing and assessing the compatibility between the conditions described in user policies and in data access requests. In the ODRL-based system proposed in this Thesis, this process involves comparing the data subjects' *odrl:Offer*, stored in their Pods, with an *odrl:Request* of a data requester, app, service or agent. As such, when considering two sets of concepts representing an offer and a request, the matching algorithm may employ two distinct and incompatible approaches to determining access to data. The first approach, the one proposed in this Thesis to cater to the specific requirements of GDPR consent described in Chapter 5, which is also the more commonly used semantically, involves treating classes as sets and determining access based on set membership. In this approach, if a class P is a superclass of C, a request for accessing P would also allow access to C because every member of C is inherently a member of P. However, a request for accessing C would not grant access to P since not all members of P are necessarily members of C. This method has also been previously utilised in matching policies for GDPR compliance by Bonatti et al. [2020].

The second approach is based on determining the applicability of a concept according to its specificity. In this method, when considering a class P and its subclass C, a request for accessing P would not extend access to C since C is more specific. Conversely, a request for accessing C would grant access to P as C is less specific. Employing subsumption as a criterion, in the first approach, access is granted when the user offer subsumes the data request, whereas, in the second

approach, access is granted when the data request subsumes the user offer. Hence, both mentioned approaches can be adapted in a decentralised data access ecosystem by reversing the direction of the subsumption in the policy matching algorithm.

Another factor to consider for the matching algorithm involves resolving permissions and prohibitions in terms of their evaluation order and potential conflicts. Policies can be interpreted in various incompatible ways, such as prioritising permissions and granting access upon the first one that is fulfilled – a permissive model. Contrarily, prioritising prohibitions and denying access upon the first fulfilled prohibition is considered a prohibitive model. In cases where both a permission and a prohibition apply to the same data or resource, conflict resolution is based on the prevalence of one over the other. In decentralised data environments such as Solid, the matching algorithm follows a prohibitive model, where prohibitions outweigh permissions. This means that if a request either fails to satisfy a permission or satisfies a prohibition, data access is not granted. As such, compatibility between offers and requests is only achieved when all permissions are satisfied and all prohibitions remain unsatisfied.

In light of these considerations, the policy matching algorithm described in this Thesis involves examining subsumption or satisfiability between instances of `odrl:Offer` and `odrl:Request`. The algorithm essentially verifies whether the conditions outlined in the user offer are met by the data request policy in the case of permissions, or breached in the case of prohibitions. If any prohibitions are identified, it indicates that certain conditions of the proposed data request are incompatible with the policies set by the user to govern the access to their personal data. Conversely, if no prohibitions are found and all permissions are met, the conditions are deemed compatible and access to data can be provided. In this context, Algorithm 6.1 offers pseudo-code outlining the steps of the proposed policy matching process.

The proposed algorithm mirrors the previously described prohibitive approach to matching, where the prohibitions outlined in the user offer are examined and ensured to be met before any permissions are considered. The denial of the access request occurs during prohibition checking if any of the following constraints in the user offer are found to be incompatible with the data request:

1. offer target has a data type matching ($\cap \neq \varnothing$) the target in the data request;
2. offer assignee matches[2] ($\equiv$) the assignee of the data request;
3. offer action has an access mode matching ($\cap \neq \varnothing$) the action in the data request;
4. offer has a purpose matching ($\cap \neq \varnothing$) the purpose in the data request;
5. offer has a recipient matching ($\cap \neq \varnothing$) the recipient in the data request;
6. offer has a legal basis matching ($\cap \neq \varnothing$) the legal basis in the data request;
7. offer has a technical and organisational measure matching ($\cap \neq \varnothing$) the technical and organisational measure in the data request;

[2]Representing permissions and prohibitions of intricate legal entities such as subsidiaries or company groups accurately is not feasible using equality ($=$) or subset ($\subseteq$) relations. Hence, in this Thesis, the equivalence relation ($\equiv$) is used to signify that the data requester must adhere to the legal interpretation of equality – defining this equality is beyond the scope of this Thesis.

Algorithm 6.1 Pseudo-code of the proposed OAC-based matching algorithm.

for $prohibition \leftarrow odrl{:}Offer$ **do**
  if $offer{:}target \cap request{:}target \neq \varnothing$ **then** $decision \leftarrow DENY$
  if $odrl{:}assignee \in offer{:}prohibition$ **then**
    if $offer{:}assignee \equiv request{:}assignee$ **then** $decision \leftarrow DENY$
  if $odrl{:}action \in offer{:}prohibition$ **then**
    if $offer{:}action \cap request{:}action \neq \varnothing$ **then** $decision \leftarrow DENY$
  for $constraint \leftarrow prohibition$ **do**
    if $oac{:}Purpose \leftarrow constraint$ **then**
      if $offer{:}Purpose \cap request{:}Purpose \neq \varnothing$ **then** $decision \leftarrow DENY$
    else if $oac{:}Recipient \leftarrow constraint$ **then**
      if $offer{:}Recipient \cap request{:}Recipient \neq \varnothing$ **then** $decision \leftarrow DENY$
    else if $oac{:}LegalBasis \leftarrow constraint$ **then**
      if $offer{:}LegalBasis \cap request{:}LegalBasis \neq \varnothing$ **then** $decision \leftarrow DENY$
    else if $oac{:}TOM \leftarrow constraint$ **then**
      if $offer{:}TOM \cap request{:}TOM \neq \varnothing$ **then** $decision \leftarrow DENY$
    else if $oac{:}Technology \leftarrow constraint$ **then**
      if $offer{:}Technology \cap request{:}Technology \neq \varnothing$ **then** $decision \leftarrow DENY$
    else if $oac{:}IdP \leftarrow constraint$ **then**
      if $offer{:}IdP \cap request{:}IdP \neq \varnothing$ **then** $decision \leftarrow DENY$
for $permission \leftarrow odrl{:}Offer$ **do**
  if $offer{:}target \cap request{:}target = \varnothing$ **then** $decision \leftarrow DENY$
  if $odrl{:}assignee \in offer{:}permission$ **then**
    if $offer{:}assignee \notin request{:}assignee$ **then** $decision \leftarrow DENY$
  if $odrl{:}action \in offer{:}permission$ **then**
    if $offer{:}action \cap request{:}action = \varnothing$ **then** $decision \leftarrow DENY$
  for $constraint \leftarrow permission$ **do**
    if $oac{:}Purpose \leftarrow constraint$ **then**
      if $request{:}Purpose \notin offer{:}Purpose$ **then** $decision \leftarrow DENY$
    else if $oac{:}Recipient \leftarrow constraint$ **then**
      if $request{:}Recipient \notin offer{:}Recipient$ **then** $decision \leftarrow DENY$
    else if $oac{:}LegalBasis \leftarrow constraint$ **then**
      if $request{:}LegalBasis \notin offer{:}LegalBasis$ **then** $decision \leftarrow DENY$
    else if $oac{:}TOM \leftarrow constraint$ **then**
      if $request{:}TOM \notin offer{:}TOM$ **then** $decision \leftarrow DENY$
    else if $oac{:}Technology \leftarrow constraint$ **then**
      if $request{:}Technology \notin offer{:}Technology$ **then** $decision \leftarrow DENY$
    else if $oac{:}IdP \leftarrow constraint$ **then**
      if $request{:}IdP \notin offer{:}IdP$ **then** $decision \leftarrow DENY$
if $\nexists DENY$ **then** $decision \leftarrow GRANT$

8. offer has a technology matching ($\cap \neq \varnothing$) the technology in the data request; and
9. offer has an identity provider matching ($\cap \neq \varnothing$) the identity provider in the data request.

If no prohibitions are identified, the next step is to verify the permissions. The access request will be denied during permission checking if any of the following constraints in the offer are incompatible with the data request:

1. request target does not have a data type matching ($\cap = \varnothing$) the target in the offer;
2. offer assignee does not match ($\notin$) the assignee of the data request;
3. request action does not have an access mode matching ($\cap = \varnothing$) the action in the offer;
4. request purpose is not compatible or a subset ($\notin$) of the offer purpose, e.g., DPV's `ResearchAndDevelopment` in a data request does not match DPV's `AcademicResearch` purpose in an offer as `ResearchAndDevelopment` is a superclass of `AcademicResearch` and, as such, less specific;
5. request recipient is not compatible or a subset ($\notin$) of the offer recipient;
6. offer legal basis does not match ($\notin$) the legal basis of the data request;
7. request technical and organisational measure is not compatible or a subset ($\notin$) of the offer technical and organisational measure;
8. request technology is not compatible or a subset ($\notin$) of the offer technology; and
9. request identity provider is not compatible or a subset ($\notin$) of the offer identity provider.

The described procedures are applied to all permissions and prohibitions outlined in the data subject's offer. If all permissions and prohibitions are met without any violations, access to the data can be authorised.

Table 6.1 showcases a set of data access agreement's outcomes illustrating the functioning of the matching algorithm concerning permissions and prohibitions, focusing on data type and purpose constraints. In a semantic-based architectural design, evaluating equivalence ($\equiv$), intersection ($\cap$), and subset ($\subseteq$) necessitates additional considerations beyond the mere interpretation of `owl:sameAs` or `rdfs:subClassOf` properties. For instance, comparing *Academic Research* as a purpose with a data request for *Research and Development* purpose, using subset ($\subseteq$) for permissions or intersection ($\cap$) for prohibition, mandates both purposes to be articulated in a manner enabling such 'hierarchical' or 'set-based' interpretations. In such a case, the matching algorithm entails interpreting *Academic Research* as a *narrower concept* or a *subset* of *Research and Development*, a relationship that can be denoted through various semantic properties, from distinct vocabularies, such as `rdfs:subClassOf`, `skos:broader`, `dcterms:isPartOf`, or even an ad-hoc property such as `ex:specialisationOf`. Further complexity emerges when considering the compatibility of purposes, as such relationships cannot be specified in a hierarchical manner.

Hence, any implementation of an OAC-based, policy matching algorithm must be aware of such relationships and carefully consider when it makes sense to employ equivalence, intersection, and subset methodologies using established Semantic Web interpretations such as the `rdf:type`

Table 6.1: Examples of outcomes of the policy matching algorithm.

Offer			Request		Outcome	
Rule	Purpose	Data	Purpose	Data	Decision	Reason
Prohibition	Academic research	Contact	Research and development	Age	DENY	request purpose ∩ offer purpose ≠ ∅
Prohibition	Academic research	Age range	Payment	Age	DENY	request data ∩ offer data ≠ ∅
Prohibition	Academic research	Contact	Payment	Age	GRANT	request purpose ∩ offer purpose = ∅ request data ∩ offer data = ∅
Permission	Academic research	Age	Commercial research	Age	DENY	request purpose ∉ offer purpose
Permission	Research and development	Age	Academic research	Age range	GRANT	request purpose ⊆ offer purpose request data ∩ offer data ≠ ∅

and `rdfs:subClassOf` properties. As such, to facilitate the consistent application and interpretation of the algorithm, a standardised specification of vocabulary terms is essential. This specification should clarify how concepts are expressed and how they are to be interpreted within the policy matching process. For instance, it should specify that any purpose term in an offer or request policy *MUST* be an instance of `Purpose` and *MUST* be associated with at least one concept in the purpose taxonomy using `rdf:type` or `rdfs:subClassOf` properties. By adhering to such a standardised specification, the matching algorithm can rely on these assertions to accurately interpret the constraints included in both offer and request policies. To achieve this, in this Thesis, the adoption of DPV's taxonomies is strongly encouraged when defining both offer and request terms to ensure the accuracy and explainability of the desired outcomes in the policy matching algorithm. Moreover, one of the real advantages of using the Semantic Web is that the set of vocabularies employed in such a matching algorithm can be changed or expanded without having to change the code base. However, it needs to be clear who has the authority to admit the usage of new vocabularies or extensions to existing ones, which in this particular case should reside in the data subjects. Nonetheless, data subjects may not have the expertise to make such decisions, and as such they can rely on public bodies' guidelines to make their choices or even rely on a DGA-trusted data intermediary to provide them with this vocabulary recommendation service. Finally, it should also be noted that such vocabulary recommendations could also be made in the form of mappings to other vocabularies, i.e., map DPV's purposes with the purposes defined in a specific health-data sharing ontology such as DUO.

6.2.4 Development of a 'Right of Access' API

This Section features the development of an API, which builds on the previous work on policies (Section 4.2) and rights exercising metadata (Section 4.4), to assist data controllers in the automation or simplification of the process to answer to a data subject's right of access request. Additionally, the implemented API method is utilised to implement a PoC Solid-based application whose main goal is to aid data subjects in exercising their GDPR right of access related to data stored in Solid Pods.

Automating the response to GDPR's right of access As mentioned in Section 1.3.4, the GDPR gave Web users additional personal data-related rights and data controllers the duty to fulfil

them. Given this, data controllers benefit from having the necessary information that they have to supply to data subjects in a structured format, in a way that facilitates automated responses to such right-related requests. The policy-based algorithms developed in this Chapter play an important role here, as they have as outcomes the storage of policies and logs, e.g., in Solid Pods, which can be used to accelerate these responses. Specifically, the 'Right of Access' is increasingly challenging for data controllers, as they are required not only to disclose the purposes for which the data was collected and used, or the specific types of data that were processed but also to furnish a copy of the data in question. Since this process is typically manual, the delay can have adverse effects on data subjects.

While there are API-based solutions to support the exercise of GDPR's right of access, such as the Microsoft Graph compliance and privacy APIs [2022h], Oracle's Data Privacy API [2021a] or the AppsFlyer [2022e], they mainly focus on the providing a copy of the data aspect of the right and do not provide detailed information concerning the purposes for which the data was processed, the duration of the processing and so on. As such, this metadata can be provided by the OAC-based policies proposed in this Thesis to provide a fully, legally-aligned right of access API to be used in decentralised systems such as Solid. The methodology used to develop the API encompassed the following steps:

1. An evaluation of current gaps on the right of access APIs was performed.
2. Similar regulation from other jurisdictions was reviewed in order to understand if new requirements needed to be added into consideration.
3. Semantic Web-based policies and provenance metadata, i.e., were used to understand the types of personal data being accessed, as well as the access conditions of said data.
4. The API method and documentation were developed.
5. Solid's personal data storage ecosystem was then chosen to verify the applicability of the API method as it is based on Web standards.

Beyond the identity of the data subject, the developed API includes parameters to specify what data they want to access as well as the purpose of said access to provide a more fine-grained right of access since not all data subjects will be interested in accessing all of their data at once. For example, the data subject might be interested in accessing only certain categories of health data or only accessing data used for research purposes. In this context, the primary role of the API is to retrieve the data stored, i.e., in the Solid Pod, and deliver it to the user in the form of a JSON file comprising two elements: a boolean variable that indicates whether personal data matching the request is present in the Pod, and a JSON object containing the corresponding list of identified resources.

The identity of the data subjects is verified when they log into their Pod. Once the data subject is logged in, the API can process the right of access request, starting by collecting the URI of all resources that have been accessed. This can be checked by looking at data agreement policies stored in the Pod. If the right exercise is performed without identifying other parameters, then all accessed resources present in the Pod will be returned, regardless of the type of data they contain or the purpose for which it was accessed. If the data subject is looking for a specific set of personal data categories, then this request should be passed as a parameter so that only those categories are

returned. The same exercise can be done if the data subject only wants to access data used for a specific purpose. Both these use cases can be implemented relying on the proposed OAC-based policies as they contain the information on what data is being access and what purpose it is being accessed for.

Exercising the right of access to Solid Pod data As data subjects require the ability to specify particular types of personal data and/or distinct purposes for data access, to be able to exercise a more detailed right of access, the developed Solid application integrates two drop-down trees. These trees utilise DPV's taxonomies of personal data and purposes to structure their content. Figure 6.6 provides a glimpse of these trees. The chosen categories are then utilised to populate the API request. An illustration of a response to a right of access request, along with the corresponding information retrieved, is depicted on (the right side of) Figure 6.6. For each resource that is retrieved, the URI, the category of personal data contained in the file, the entities that accessed the data, and a list of the policies governing access to the resource are provided. Additionally, a download button is included to enable data subjects to obtain a copy of the resource data. This approach represents an advancement over the current solutions by offering detailed information regarding the personal data categories present in the data, along with information on how it was used, such as the purposes for which they were accessed, in addition to facilitating the provision of a data copy. Moreover, right exercise metadata is kept in the Pod using the work proposed in Section 4.4.

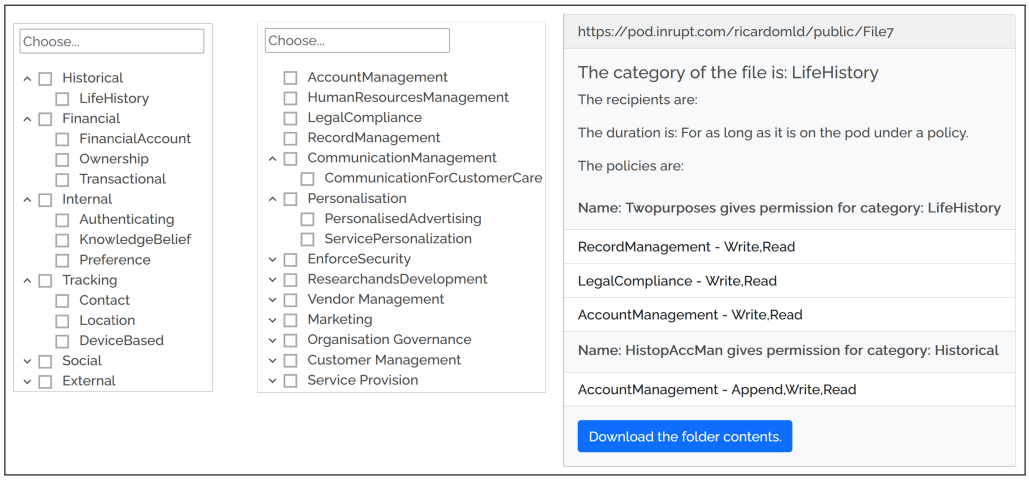


Figure 6.6: Screenshot of an example Solid-based application that uses the implemented Right of Access API.

Maintenance, preservation, and future work The publication and archival of the developed work are performed according to the methodology described in Section 3.7.3. Furthermore, the source code is hosted at https://w3id.org/people/besteves/access-right/api and https://w3id.org/people/besteves/access-right/solid (for the API and for the Solid app, respectively), under the MIT license. These repositories can also be utilised by their users to suggest new features to be added to the work, as well as to document bugs through GitHub Issues.

In future developments, enhancements to the API could offer a more comprehensive response to a right of access request. This would entail providing information on other data subject rights and clear details regarding data recipients, including their identities and contact information. Moreover, certain practical considerations have been overlooked. For instance, it has been assumed that resources stored in the Pod under the influence of a policy can be automatically accessed through the API during their storage period. This functionality could be expanded to accommodate new policy constraints, such as time-limited storage or periodicity constraints. Additionally, new parameters for filtering requested data could be incorporated into both the API and the Solid application.

6.3 Proof of concept implementation for health data sharing

In this Section, a proof of concept implementation of the proposed algorithms for offer instantiation and policy matching towards the generation of a data access agreement is described for a specific use case involving health data sharing.

6.3.1 Background and motivation

In this Thesis, a health data sharing use case was selected to showcase the strengths of the proposed algorithm since the exchange of health-related data presents significant potential for advancing research and leveraging advanced computational and statistical techniques to drive progress in healthcare. However, due to its sensitive nature and potentially significant impact if misused, the sharing and utilisation of health-related data are highly regulated at both legal and institutional levels, e.g., *"data concerning health"* is a special category of personal data under the GDPR, i.e., Article 9.1 [2016b], and as such its processing is prohibited unless one of the legal grounds of Article 9.2 applies.

Currently, institutions such as hospitals handle each health data request through a dedicated committee tasked with evaluating and making decisions regarding the release of such data under their care. To facilitate this process, the Global Alliance for Genomics and Health[3] (GA4GH) was established as an international consortium focused on the development of standards and the promotion of responsible sharing of genomics and health data. Among its various resources, aimed at different aspects and processes of health-related data sharing, GA4GH has introduced a machine-readable ontology known as the Data Use Ontology[4] (DUO). DUO [Lawson et al., 2021, Rehm et al., 2021] was designed to express Data Use Limitations (DULs), i.e., conditions and constraints defined by data providers which should be respected by data requesters to use said data.

DUO is an OWL ontology aligned with the Open Biological and Biomedical Ontology[5] (OBO). By utilising OBO's upper level ontologies, DUO ensures semantic interoperability with a range of biomedical ontologies belonging to the OBO family of ontologies. As such, DUO's main purpose is related to annotating datasets with DUL codes to specify usage conditions, articulate data usage

[3] https://www.ga4gh.org/ (accessed on 2 April 2024)
[4] http://purl.obolibrary.org/obo/duo (accessed on 2 April 2024)
[5] https://obofoundry.org/ (accessed on 2 April 2024)

requests, and automatically identify or discover compatible datasets by comparing requests with datasets' usage conditions.

DUO and existing efforts DUO concepts are organised into three taxonomies:

1. The 'Data Use Permission' taxonomy, denoted by the base class obo:DUO_0000001, encompasses permissions related to purposes for data usage.
2. The 'Data Use Modifier' taxonomy, denoted by the base class obo:DUO_0000017, covers additional conditions to be applied alongside data use permissions.
3. The 'Investigation' taxonomy, denoted by the base class obo:OBI_0000066 from the Ontology for Biomedical Investigations[6] (OBI), delineates planned conditions for which data is being requested.

Furthermore, the obo:DUO_0000010 property implements a *'is restricted to'* relation, used to limit certain concepts to certain contexts, e.g., to restrict the obo:DUO_0000022 concept, which indicates usage permitted within a geographic region, to terms from obo:GAZ_00000448, the base concept from the Gazetteer (places) ontology[7].

DUO originated from prior endeavours aimed at establishing consent codes for data usage and leveraging them as machine-readable information for automated access to data. Its initial development drew upon the Consent Codes by Dyke et al. [2016], which specified concepts for data usage based on consent permissions. Subsequently, DUO incorporated a few terms from the Automatable Discovery and Access Matrix (ADA-M) framework [Woolley et al., 2018], which shares akin objectives and concepts. As such, the intended utilisation of DUO revolves around facilitating the recording of consent for sharing and reusing biomedical datasets, as outlined by Lawson et al. [2021] and Rehm et al. [2021], the latter highlighting the usage of DUO in distinct GA4GH initiatives.

The Data Use Oversight System[8] (DUOS) is a platform built upon DUO aimed at facilitating semi-automated management of access health-related datasets. It leverages DUO annotations to incorporate new datasets and data access requests, which are subsequently matched using an algorithm based on hierarchical compatibility. Figure 6.7 illustrates the algorithm followed by this platform, where the 'Data Use Codes' correspond to concepts from the previously mentioned DUO 'Data Use Permission' taxonomy, the 'Data Use Requirements' to the 'Data Use Modifier' taxonomy, and the 'Research Purpose Terms/Data Use Categories' to the 'Investigation' taxonomy. This algorithm entails matching the dataset's access conditions with the data requester's conditions by relying on the subclass relations between them. The results of this matching algorithm are then reviewed by a 'Data Access Committee' to finalise the conditions of the access agreement and provide the data requested with access to the relevant datasets, comparable with the decision-making processes of human data access committees [Cabili et al., 2021]. The NHGRI Genomic Data Science Analysis, Visualization, and Informatics Lab-space project[9] implemented a large-scale pilot using the DUOS platform implementation [Schatz et al., 2022].

[6] http://obi-ontology.org/ (accessed on 3 April 2024)
[7] https://environmentontology.github.io/gaz/ (accessed on 3 April 2024)
[8] https://duos.broadinstitute.org/ (accessed on 2 April 2024)
[9] https://anvilproject.org (accessed on 3 April 2024)

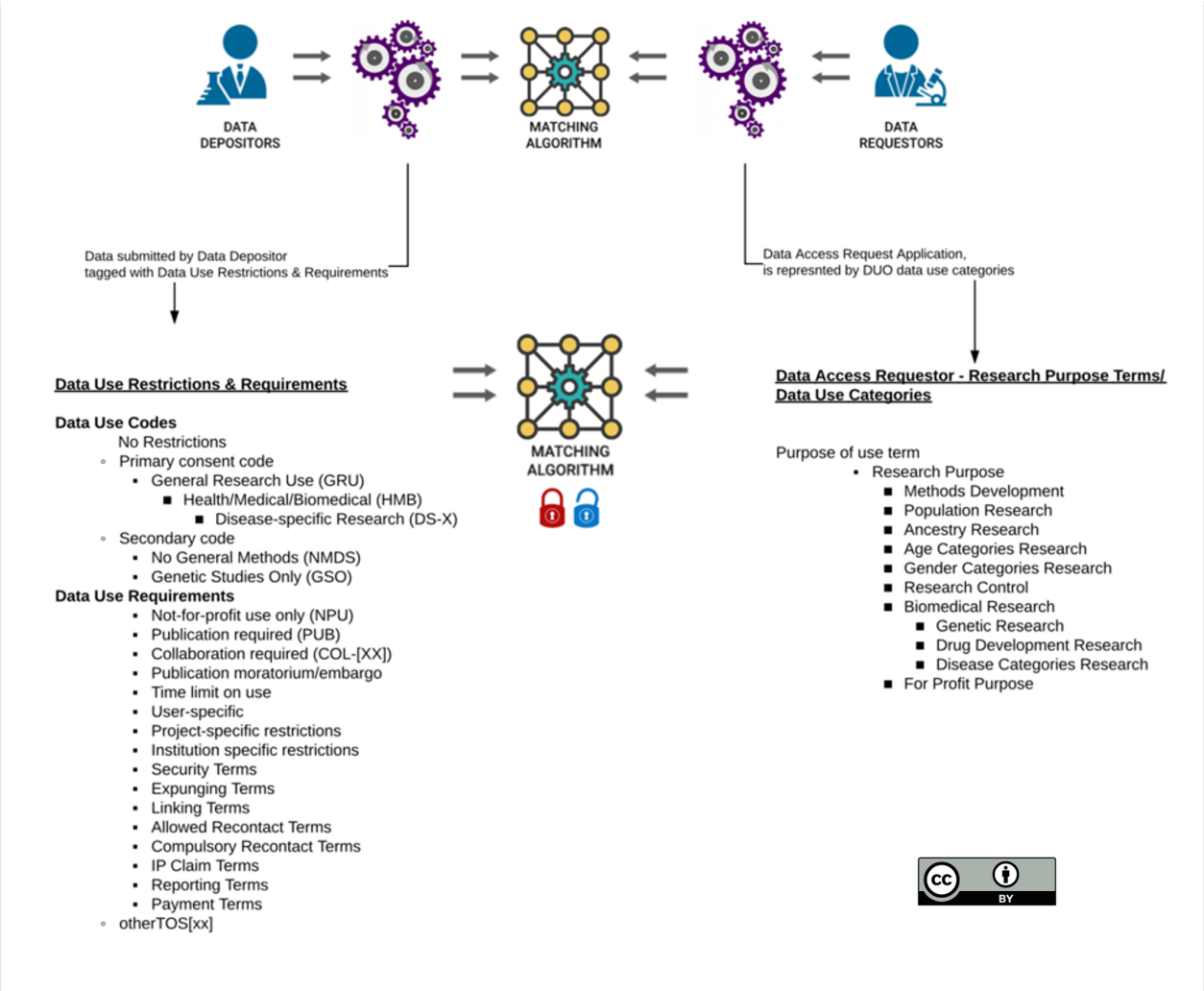


Figure 6.7: DUO-based matching algorithm, adapted from https://github.com/EBISPOT/DUO.

Beyond DUOS, the DUO specification has been integrated into various other works, including a specification of informed consent for health and genomics research in Africa [Nembaware et al., 2019], a blockchain-based consent model for health data sharing [Jaiman and Urovi, 2020], and an online platform that offers dynamic consent interfaces and tools for large-scale genomics research programs [Haas et al., 2021]. Furthermore, DUO is mentioned in the Data Tags Suite (DATS), where it is considered a candidate vocabulary within its framework for discovering metadata-based data access conditions [Alter et al., 2020]. Additionally, it plays a role in the roadmap of European infrastructures for accessing a large number of human genomes [Saunders et al., 2019]. Moreover, Amith et al. [2022] demonstrate the usage of DUO for the representation of consent metadata, employing SWRL[10] to execute permissive rules, and Grabus and Greenberg [2019] provide a comprehensive overview of rights and licensing initiatives, approaches, and tools for health data sharing, including DUO.

[10] https://www.w3.org/Submission/SWRL/ (accessed on 3 April 2024)

Challenges of the DUO specification DUO expresses DULs as concepts with human-readable definitions, utilising the `obo:IAO_0000115` property, e.g., as `skos:definition` is used to define SKOS concepts. This limits their utility to humans or machines that operate solely on known concepts. Furthermore, DUO concepts lack linkage to relevant legal concepts, leading to ambiguity regarding the implications of their usage in strongly legislated jurisdictions like the EU, where the GDPR [2016b] introduces additional accountability and compliance requirements that must be acknowledged and adhered to. While existing documentation mentions that the applicability of laws falls under the responsibility of the adopter and that DUO terms have not been evaluated for GDPR compliance, it is crucial for data subjects and data controllers to ensure compatibility with existing regulations. As such, the absence of such support from the DUO specification poses a risk of hindering interoperability as additional approaches need to be taken to fulfil legal requirements. Moreover, with the EU push to have a common 'Health Data Space'[11], machine-readability and automation will play a pivotal role in facilitating legally-aligned health data exchange.

Furthermore, no instructions were found on how to associate DUO conditions with datasets nor on how the data access agreements matching algorithm should work. While the DUOS framework provides a comprehensive description of how DUO can be utilised, it lacks detailed guidance on the matching algorithm. DATS [Alter et al., 2020] also acknowledges the challenge of using DUO for defining permissions and prohibitions, suggesting ODRL as an alternative model for clearer articulation of permissions and prohibitions.

Proposed improvements over the DUO specification To achieve *true machine-readability*, DUO concepts must be represented with permissions, prohibitions, constraints, and duties to form machine-readable rules, by leveraging semantic standards for the expression of asset usage conditions. By formalising the DULs embedded within the descriptions of each DUO concept as a set of rules, these become explicit and can be attached and sent alongside the data for future inspection.

To assess the compatibility of a data request with the dataset's DULs, both the conditions set by the data provider and those articulated by the data requester for data use should be formulated as policies. These policies can then be matched to determine if the intended use aligns with the dataset's conditions, using the same approach as the one presented in Section 6.2. While DUO is currently being utilised in this manner, as evidenced in systems like DUOS, the matching relies on hierarchical compatibility between data request and data use conditions established through a subclass relationship, i.e., a data request DR is a subclass/superclass of a data use condition DUC. This approach has limitations in terms of its capacity and expressiveness for delineating fine-grained rules to utilise in automated systems, as not all relevant pieces of information can be explicitly captured in distinct concepts. For instance, DUO's `DUO_0000006` indicates through a sole human-readable label that use is allowed for health/medical/biomedical research purposes, not including the study of population origins or ancestry, while the same information can be much more explicitly declared using ODRL policies with permission and prohibition rules with purpose constraints.

[11] https://ec.europa.eu/health/ehealth-digital-health-and-care/european-health-data-space_en (accessed on 2 April 2024)

More significantly, in order to automate the generation of data access/usage agreements effectively, a set of criteria should be considered when selecting a vocabulary to articulate such conditions:

(i) the level of expressiveness to define rules and policies, encompassing the ability to express actions, purposes, or other constraints as distinct concepts that can be autonomously specified and evaluated, and combined in distinct ways to represent various types of policies;

(ii) the capability to associate and verify their conformity and adherence to legal requirements, e.g., such as the GDPR; and

(iii) the capacity to specify access/usage conditions in a machine-readable format and utilise them for assessing the accuracy and comprehensiveness of information that should be in such a data agreement.

With the aforementioned motivation in mind, this Thesis proposes an approach to explicitly represent DUO concepts using RDF, leveraging ODRL and the efforts declared in previous Sections related to an OAC-based architecture for decentralised access to data. The choice of using ODRL, beyond being a W3C Recommendation for the expression of policies, is supported by the following motives:

(i) it is RDF-based, ensuring machine-readability;

(ii) it encompasses concepts that model domain-specific and legally relevant terms to depict constraints, such as spatial and temporal operators, as well as support various types of policies like offers, requests, and agreements;

(iii) its usage can be validated, and efforts are underway within the W3C ODRL CG to actively develop a formal semantics specification [Fornara et al., 2023];

(iv) it facilitates the development of extensions through ODRL profiles, offering the flexibility to tailor ODRL to specific requirements – as proved by this Thesis' work on OAC (Section 4.2), which connects ODRL with legal requirements using DPV and can be extended to cater for legal requirements of health data sharing; and

(v) backward compatibility can be ensured – existing DUO-based systems can adopt the practices suggested in this Section and continue being compatible with the DUO specification. Also, DUO users can select which aspects of this solution they want to incorporate into their system.

As such, based on the algorithms described in Section 6.2, in this Section, (i) DUO concepts are modelled as ODRL policies, (ii) such policies are instantiated as ODRL offers for access to health-related datasets, (iii) offers are matched with incoming data requests to generate permissive or prohibitive data agreements, and (iv) work on OAC is recycled to deal with GDPR obligations for the processing of health data.

6.3.2 Re-modelling DUO concepts with ODRL

As outlined in the previous Section, DUO concepts are organised into three taxonomies with textual annotations representing the access conditions. This Section's main goal is to examine this implicit information and articulate it explicitly as ODRL policies.

Additionally, it aims to ensure compatibility with current and future GA4GH activities, minimising significant disruptions to existing workflows that already use DUO. Therefore, the role of ODRL is not to supplant DUO but rather to supplement it by providing extra machine-readable information for each DUO concept. This additional information makes explicit the conditions currently embedded in DUO's textual annotations, enabling them to be validated, matched, and used for data access in an automated fashion. An ODRL-improved DUO specification can also be used for additional information duties, such as record-keeping by institutions or other legal tasks. Following the algorithm proposed in Section 6.2, DUO's taxonomies are modelled as ODRL offers and requests and used as inputs of the policy matching algorithm to generate data access agreements.

Identifying ODRL equivalents for DUO concepts For each DUO term, the initial task involved discerning the rules and constraints it encompasses, through the analysis of its textual description, to determine whether it should be modelled as a permission, prohibition, or obligation, as well as understanding the specific context in which they should be applied. With this analysis, the redundancy and overlap between DUO's data use permissions and modifiers is noticeable, as both include purpose-based policies without a clear differentiation in their semantics. For instance, DUO_0000011 denotes permission, while DUO_0000044 signifies prohibition, both for 'population origins or ancestry research' purposes, with the former categorised as a data use permission and the latter a modifier.

Hence, in this Thesis, the restructuring of the DUO taxonomies is proposed by introducing a unique purpose-based taxonomy that models research concepts as ODRL policies, with variants for permissions and prohibitions, and duties applied over permissions with specific research purposes. For instance, the DUO concept for code HMB, which states that "[t]his data use permission indicates that use is allowed for health / medical / biomedical purposes (HMB); [and] does not include the study of population origins or ancestry" in its textual definition, should be articulated as being a permission for a purpose of type HMB and not a purpose of type POA. This approach allows for a more precise expression and application of DUO's HMB concept by revealing its underlying concepts (purpose) and the rules governing it (permission).

Furthermore, upon analysing DUO's concepts and their inherent access conditions, ODRL rules were modelled to represent the conditions of each DUO concept. In cases where ODRL lacks the necessary concepts, an extension to its existing concepts is proposed to cover the missing ones. In this context, for each concept, an `odrl:Set` instance was modelled to represent the DUO concept's intrinsic rules. These policies are then amalgamated into an `odrl:Offer`, to serve as a unified policy for a dataset, as detailed in Section 6.2.2 regarding the instantiation of offer policies. A comprehensive compilation of the performed interpretations for each DUO concept is provided in Table 6.2. As will also be mentioned later in this Section, the original DUO concept definitions will still be kept in the interpreted ODRL policies using the `rdfs:label` property. Moreover, due to the interpretive nature of this analysis, difficulties were found in interpreting specific expressions such as 'is limited to', a term that suggests that usage is restricted solely within a particular scope. If this interpretation holds true, DUO developers should clarify how potential conflicts arising from rules expressing exclusive limitations versus other permissive expressions, such as 'is allowed for', should be resolved. Hence, in this Thesis, ODRL's capabilities are used to articulate these requirements as permissions, prohibitions, and obligations, and, subsequently,

existing works on ODRL [Pellegrini et al., 2018a, De Vos et al., 2019] and OWL [Bonatti et al., 2020] reasoners can be employed to aid solving conflict resolution issues.

Presently, DUO concepts are confined to delineating conditions for data use, suggesting the usage of external ontologies to incorporate additional concepts needed for refining the scope of the term, e.g., DUO_0000007 denotes permission for disease-specific research and recommends the usage of the Mondo Disease Ontology ontology[12] for specifying the diseases to which the permission applies. Other specific concepts mentioned in DUO's textual annotations, which are not explicitly modelled, include codes inherited from previous iterations of DUO, such as purpose terms like *CC* for Clinical Care Use, or *GRU* for General Research Use. To express ODRL policies, these concepts need an explicit, individual semantic specification, so that they can be used to constrain permissions or prohibitions. Thus, in this Thesis' implementation, missing terms were identified and compiled into an ad-hoc vocabulary (DUODRL[13]) to enable the correct expression of ODRL policies for each DUO concept. This approach also offers the benefit of delivering more comprehensive documentation related to the information represented within DUO concepts. For instance, by modelling *IRB* as an ODRL policy representing the duty to get an Ethics Review Board approval, instead of just using the DUO concept, it becomes feasible to include details about the processes and requirements necessary for such reviews. It also allows additional constraints related to ethics approvals to be semantically linked with this foundational concept, such as indicating that it must be conducted before data access is provided, periodically, or before any outcomes of the research are published.

In the case of data requests, the so-called DUO 'Investigations', redundant concepts were found in relation to both the data use permissions and modifiers concepts, e.g., DUO_0000040 denotes a data request for research on a specific disease, while DUO_0000007 signifies permission for research on a specific disease. In essence, both concepts convey the same idea concerning 'research on a specific disease', albeit with one indicating a data request and the other representing the dataset's permissive conditions. As such, as previously suggested in relation to the reorganisation of DUO's 'Data Use Permission' and 'Data Use Modifier' taxonomies to be research purpose-based, a similar approach should be adopted for data requests to ensure consistency in the concepts used for requesting data. This alignment promotes clarity, reduces ambiguity, and facilitates the matching process, as the same rules can be associated to a dataset through an `odrl:Offer` and to a data request through an `odrl:Request`.

Besides specifying requirements for data access and requests to access said data, DUO concepts should also be utilised to document the result of the matching algorithm, and whether access is to be granted or not. This aspect represents an important yet unexplored domain in the current identified applications of DUO, particularly because any data sharing typically entails accompanying information about the involved parties, the provenance linked with the granting process, and details regarding how the data access conditions were satisfied at the time or what are future duties for the access to be granted. Hence, the work on this Thesis related to the generation of data access agreements, described in detail in Section 6.2.3, will be beneficial in portraying this information, not to mention the work on OAC and PLASMA, which can encompass all the aforementioned policy and metadata-related details. Furthermore, these agreements can be

[12] https://obofoundry.org/ontology/mondo (accessed on 4 April 2024)
[13] DUODRL is published at https://w3id.org/duodrl.

Table 6.2: Interpretation of DUO concepts' textual descriptions as ODRL policies.

Concept	Code	Rule Type	Constraint	Placeholder
DUO0000001	**Data Use Permission**			
DUO0000042	GRU	Permission	Purpose is :GRU	
DUO0000006	HMB	Permission	Purpose is :HMB and not :POA	
DUO0000007	DS	Permission	Purpose is :DS and mondo:0000001	:TemplateDisease
DUO0000004	NRES	Permission	Purpose is odrl:Purpose	
DUO0000011	POA	Permission	Purpose is :POA	
DUO0000011	POA	Prohibition	Purpose is not :POA	
DUO0000017	**Data Use Modifier**			
DUO0000043	CC	Permission	Purpose is :CC	
DUO0000020	COL	Duty	Action is :CollaborateWithStudyPI	
DUO0000021	IRB	Duty	Action is :ProvideEthicalApproval	
DUO0000016	GSO	Permission	Purpose is :GS or :GSG	
DUO0000016	GSO	Prohibition	Purpose is :GS and not :GSG	
DUO0000022	GS	Permission	Spatial is equal to specified :Location	:TemplateLocation
DUO0000022	GS	Prohibition	Spatial is not equal to specified :Location	:TemplateLocation
DUO0000028	IS	Permission	Assignee is :ApprovedInstitution	:TemplateInstitution
DUO0000028	IS	Prohibition	Assignee is not :ApprovedInstitution	:TemplateInstitution
DUO0000015	NMDS	Prohibition	Purpose is :MDS	
DUO0000018	NPUNCU	Permission	Assignee is :NonProfitOrganisation and Purpose is :NCU	
DUO0000018	NPUNCU	Prohibition	Assignee is :ForProfitOrganisation and Purpose is :NCU	
DUO0000018	NPUNCU	Prohibition	Assignee is :NonProfitOrganisation and Purpose is not :NCU	
DUO0000046	NCU	Permission	Purpose is :NCU	
DUO0000046	NCU	Prohibition	Purpose is not :NCU	
DUO0000045	NPU	Permission	Assignee is :NonProfitOrganisation	
DUO0000045	NPU	Prohibition	Assignee is :ForProfitOrganisation	
DUO0000044	NPOA	Prohibition	Purpose is :POA	
DUO0000027	PS	Permission	Project is :ApprovedProject	:TemplateProject
DUO0000027	PS	Prohibition	Project is not :ApprovedProject	:TemplateProject
DUO0000024	MOR	Duty	Action is odrl:distribute :ResultsOfStudies with odrl:dateTime	:TemplateDateTime
DUO0000019	PUB	Duty	Action is odrl:distribute :ResultsOfStudies	
DUO0000012	RS	Permission	Purpose is specified :Research	:TemplateResearch
DUO0000012	RS	Prohibition	Purpose is not specified :Research	:TemplateResearch
DUO0000029	RTN	Duty	Action is :ReturnDerivedOrEnrichedData	
DUO0000025	TS	Permission	Time is less than specified :TemplateDateTime	:TemplateDateTime
DUO0000026	US	Permission	Assignee is :ApprovedUser	:TemplateUser
DUO0000026	US	Prohibition	Assignee is not :ApprovedUser	:TemplateUser
OBI0000066	**Investigation**			
DUO0000034		Permission	Purpose is :AgeCategoryResearch	
DUO0000034		Permission	Age is specified :Age	:TemplateAgeCategory
DUO0000033		Permission	Purpose is :POA	
DUO0000037		Permission	Purpose is :HMB	
DUO0000040		Permission	Purpose is :DS and mondo:0000001	:TemplateDisease
DUO0000039		Permission	Purpose is :DrugDevelopment	
DUO0000038		Permission	Purpose is :GS	
DUO0000035		Permission	Purpose is :GenderCategoryResearch	
DUO0000035		Permission	Gender is specified :Gender	:TemplateGender
DUO0000031		Permission	Purpose is :MDS	
DUO0000032		Permission	Purpose is :PopulationGroupResearch	
DUO0000032		Permission	Population is specified :Population	:TemplatePopulation
DUO0000036		Permission	Purpose is :ResearchControl	

employed in automated processes that periodically verify if the pending duties on an agreement still have to be or are already fulfilled, e.g., the publication of the research results.

DUO concepts as `odrl:Set`s Considering the points raised above, every DUO restriction should be portrayed as a `odrl:Set` instance, which, according to its definition, should include at least one permission, prohibition, or obligation, along with one associated data asset, to be deemed a semantically correct ODRL policy. Its utilisation does not confer any access privileges but solely denotes a set of rules that are applicable to the resource. Considering the human-readable labels linked with each DUO concept, `odrl:permission` was used when the condition permitted access to data, `odrl:prohibition` when it prohibited access, and `odrl:duty` when it outlined obligations to allow data access. DUO's textual descriptions are still kept in the policies via the `rdfs:label` property and the connection with the original DUO concept using `skos:exactMatch`.

Constructing a valid ODRL policy proved challenging due to the necessity of specifying *"any form of identifiable resource"* [Iannella et al., 2018]. This challenge arises because DUO concepts solely represent specific access and usage requirements unrelated to a specific dataset. Additionally, DUO does not indicate how to express values associated with requirements such as research in specific diseases or temporal duration of access. To ensure the validity of ODRL policies and provide clarity on how to later apply or instantiate them as an `odrl:Offer` for a particular dataset, the base class `TemplateQuery` was introduced in DUODRL. Instances of this class serve as a placeholder to be substituted with the actual value(s) obtained by executing a SPARQL query which is linked with the placeholder through the `sparqlExpression` property. In Table 6.2, these instances are denoted in the 'Placeholder' column. Examples of this approach are illustrated in Listing 6.1 for an `odrl:Set` representing a DUO data use permission and a modifier. In this case, the placeholders are employed to indicate datasets that are not instantiated yet and a publication date that is still not determined. These placeholders can be later replaced with actual instances in real policies, e.g., by utilising SPARQL queries.

The indication of scoped restrictions, such as specifying the geographic location when use is confined to a particular region, was also not void of issues. Within DUO, the `obo:DUO_0000010` property is used to describe a 'is restricted to' relationship, which is intended to specify the specific values or instances of variables such as diseases or locations. However, the descriptions of DUO concepts only mention that *"this should be coupled with an ontology term describing the (concept) the restriction applies to"*, without providing an example illustrating how it should be utilised.

Hence, to address this challenge, four possible approaches were identified:

(i) utilising OWL class expressions[14];

(ii) employing SHACL shapes to specify a constraint;

(iii) developing a new ODRL `Operator` that accepts property paths; or

(iv) directly stating the term as an instance of the scoping concept, e.g., for disease-specific restrictions, the concept would be instantiated as both the appropriate DUO concept and the

[14] https://protegeproject.github.io/protege/class-expression-syntax/ (accessed on 6 April 2024)

disease concept).

Each of these approaches influences how a condition is articulated and impacts the performance and functionality of the policy matching algorithm, e.g., leveraging approach (i) would necessitate the execution of an OWL2 reasoner before the matching process, while approach (ii) would require a SHACL validator. In this Thesis, the fourth approach was taken by declaring the concept as an instance of both the DUO and the scoping classes. This method was chosen for its simplicity and absence of the need for additional tools or modifications to ODRL. Moreover, if no extra tools or modifications are needed and a different method in the future is found to be more effective then it can easily replace the one proposed in this Section. Additionally, by reusing the existing ODRL and OAC models, it is also possible to reuse the algorithms specified in Section 6.2.

The `odrl:Set` formulated to represent DUO's concept of *"general research use"* and the one to represent DUO's concept of *"publication moratorium"* are detailed in Listing 6.1. When instantiated, beyond additional provenance metadata such as the `dcterms:creator` and `dcterms:issued` properties to indicate the creator and date of issuance of the offer, the `duodrl:TemplateDataset` must be replaced by the dataset's identifier and, in the case of the moratorium, a date must be instantiated to express until when the distribution of the outcomes of the research study is prohibited.

Dataset policies as `odrl:Offer`s Datasets can be annotated with several DUO concepts. In this context, an instantiation algorithm must be used to collect the `odrl:Set`s that should be applied to the dataset to be instantiated into an `odrl:Offer`. For this, the algorithm for offer instantiation presented in Section 6.2.2 can be recycled for the particular use case of DUO. As such the following steps were taken to instantiate dataset policies as `odrl:Offer`s:

1. For a given dataset, retrieve all data use permissions and data use modifiers that the dataset was tagged with.
2. Filter out duplicated policies.
3. If a retrieved policy contains instances of `TemplateQuery`, e.g., `TemplateDataset`, execute their SPARQL queries to replace the template with the retrieved values.
4. For each retrieved policy, fetch relevant `odrl:Permission` and `odrl:Prohibition` rules and merge them in a single `odrl:Offer`.
5. A link to each 'original' policy is maintained in the final `odrl:Offer` by using the `dcterms:source` property.
6. Add provenance information to the `odrl:Offer`, e.g. `dcterms:issued` for when the offer was instantiated and `dcterms:creator` for the issuer of the policy.

An example `odrl:Offer`, which merges the `DUO_0000042`, `DUO_0000025` and `DUO_0000020` policies as conditions to access the `ex:EHR` dataset, is presented in Listing 6.2.

Moreover, in decentralised data environments, the discovery of the dataset's location will play an important role as during the offer instantiation algorithm the `TemplateDataset` variable must be substituted by the real location of the dataset to which the offer applies. As such, the SPARQL query identified in Listing 6.3 showcases an example of how to find the location of medical

Listing 6.1 `odrl:Sets` representing DUO_0000042, a data use permission for general research purposes (GRU), and DUO_0000024, a data use modifier to indicate that the results of the research should be published only after a moratorium period (MOR).

```
duodrl:DUO_0000042 a odrl:Set ;
    odrl:uid duodrl:DUO_0000042 ;
    odrl:profile oac: ;
    rdfs:label "DUO_0000042: This data use permission indicates that use
     ↪ is allowed for general research use for any research purpose (GRU
     ↪ - general research use)"@en ;
    skos:exactMatch obo:DUO_0000042 ;
    skos:editorialNote "We interpreted this as a permission for a
     ↪ research purpose that is an instance of general research purposes
     ↪ (GRU). We note that the DUO concept description consists of
     ↪ specific areas of research which can be additionally indicated if
     ↪ meant as an exclusive list, i.e., purpose must be one of those,
     ↪ or provide as subclasses of GRU if meant to be provided as a
     ↪ selection, i.e., purpose can be expressed as one of those"@en ;
    odrl:permission [
        odrl:action odrl:use ;
        odrl:target duodrl:TemplateDataset ;
        odrl:constraint [
            odrl:leftOperand oac:Purpose ;
            odrl:operator odrl:isA ;
            odrl:rightOperand duodrl:GRU ] ] .

duodrl:DUO_0000024 a odrl:Set ;
    odrl:uid duodrl:DUO_0000024 ;
    rdfs:label "DUO_0000024: This data use modifier indicates that
     ↪ requestor agrees not to publish results of studies until a
     ↪ specific date (MOR - publication moratorium)"@en ;
    skos:exactMatch obo:DUO_0000024 ;
    skos:editorialNote "We interpret this as a prohibition to not publish
     ↪ results until the specified date as indicated by
     ↪ TemplateStudyResultsPublicationDate placeholder. We also note
     ↪ that this rule may be expressed using odrl:delayPeriod or
     ↪ odrl:dateTime if it should be expressed as a (delayed) permission
     ↪ to publish instead of a prohibition"@en ;
    odrl:prohibition [
        odrl:target duodrl:TemplateDataset ;
        odrl:action odrl:distribute ;
        odrl:output duodrl:ResultsOfStudies ;
        odrl:constraint [
            odrl:leftOperand odrl:dateTime ;
            odrl:operator odrl:lt ;
            odrl:rightOperand duodrl:TemplateStudyResultsPublicationDate
             ↪ ] ] .
```

Listing 6.2 An example `odrl:Offer` containing a permission for general research use (GRU), from DUO_0000042, a time limit on the use (TS), from DUO_0000025, and a duty to collaborate with the studies' primary investigator (COL), defined from DUO_0000020.

```
ex:offer_for_GRU_TS_COL a odrl:Offer ;
    odrl:uid ex:offer_for_GRU_TS_COL ;
    odrl:profile oac: ;
    rdfs:label "Offer to use dataset for GRU within time limits while
     ↪ collaborating with the primary study investigator" ;
    odrl:assigner ex:provider ;
    odrl:target ex:EHR ;
    odrl:action odrl:use ;
    dcterms:source duodrl:DUO_0000042, duodrl:DUO_0000025,
     ↪ duodrl:DUO_0000020 ;
    dcterms:issued "2024-04-06"^^xsd:date ;
    odrl:permission [
        odrl:duty [ odrl:action duodrl:CollaborateWithStudyPI ] ] ;
    odrl:permission [
        odrl:constraint [
            odrl:leftOperand odrl:elapsedTime ;
            odrl:operator odrl:lteq ;
            odrl:rightOperand "2024-12-31"^^xsd:date ] ] ;
    odrl:permission [
        odrl:constraint [
            odrl:leftOperand oac:Purpose ;
            odrl:operator odrl:isA ;
            odrl:rightOperand duodrl:GRU ] ] .
```

health data datasets, tagged as such using the DPV's personal data categories extension, and in particular the `pd:MedicalHealth` term, and using the PLASMA conforming specification of a data registry. By running such a query it will be possible to find the location of specific datasets in decentralised settings such as the ones promoted by Solid. Similar queries can also be executed for decentralised environments beyond Solid or even within centralised environments.

Listing 6.3 SPARQL query to retrieve template datasets locations from a PLASMA data registry specified in a Solid Pod.

```
SELECT DISTINCT ?DatasetsLocation WHERE {
    ?DataRegistry a plasma:DataRegistry .
    ?DataRegistry dcat:catalog ?Entry .
    ?Entry a dcat:Catalog .
    ?Entry dcat:themeTaxonomy pd:MedicalHealth .
    ?Entry dcat:dataset ?DatasetsLocation .
}
```

Data access outcomes as `odrl:Agreements` To provide access to a dataset, data-handling environments must be able to match dataset policies with a specific request being made e.g. for research with health data. For this, DUO's requests must be instantiated as `odrl:Requests` to be matched with the datasets `odrl:Offers`. A request for health, medical, or biomedical

research studies (DUO_0000037) is presented in Listing 6.4.

Listing 6.4 An `odrl:Request` containing a request for health, medical, or biomedical research (HMB) created from DUO_0000037.

```
ex:request_for_HMB a odrl:Request ;
    odrl:uid ex:request_for_HMB ;
    odrl:profile oac: ;
    rdfs:comment "Request for health, medical, or biomedical research" ;
    dcterms:source duodrl:DUO_0000037 ;
    dcterms:issued "2024-05-01"^^xsd:date ;
    odrl:permission [
        odrl:action odrl:use ;
        odrl:target duodrl:TemplateDataset ;
        odrl:assignee ex:requester ;
        odrl:constraint [
            odrl:leftOperand oac:Purpose ;
            odrl:operator odrl:isA ;
            odrl:rightOperand duodrl:HMB ] ] .
```

In this context, an algorithm must be used to collect the conditions under which access to data is allowed/denied. As such, the algorithm for policy matching presented in Section 6.2.3 can be recycled for the particular use case of DUO matching for health datasets. Considering this, the following steps were taken to instantiate dataset access agreements as `odrl:Agreements`:

1. Retrieve the data requester's `odrl:Request` and the dataset's `odrl:Offer`.
2. Match the `odrl:Offer` with the `odrl:Request`.
3. Record the outcome of the matching algorithm, where the `odrl:target` property specifies the dataset to be accessed, and the `odrl:assignee` and `odrl:assigner` properties identify the data requester and the provider, respectively.
4. If the matching result is positive, i.e., the request and the offer are compatible, then access is permitted by employing a permission with constraints on the requested purpose for access, along with any additional constraints such as spatial or temporal constraints or duties on the requester. If access is denied, similar information is included in the policy as a prohibition.
5. Utilise the `dcterms:references` property to associate the agreement with the `odrl:Offer` and `odrl:Request` that were used to generate it.
6. Include provenance and other relevant information, such as the `dcterms:issued` property, to document the creation and acceptance of the agreement among the parties.

Following these steps, an agreement for access to data can be reached. An example representation of a DUO-based data access agreement, modelled as an `odrl:Agreement`, to give access to data until 2024–12–31 for the purpose of health, medical, or biomedical research studies is presented in Listing 6.5, including also a duty which must be fulfilled after data access.

As previously mentioned in the policy matching algorithm described in Section 6.2.3, the matching process described in step 2 of agreement generation operates by comparing and assessing the compatibility between the conditions described in dataset policies and in data requests. Since DUO

Listing 6.5 An `odrl:Agreement` representing a decision to access a `ex:EHR` dataset from `ex:provider`.

```
ex:agreement_for_HMB a odrl:Agreement ;
    odrl:uid ex:agreement_for_HMB ;
    odrl:profile oac: ;
    dcterms:references ex:offer_for_GRU_TS_COL, ex:request_for_HMB ;
    dcterms:issued "2024-05-01"^^xsd:date ;
    odrl:action odrl:use ;
    odrl:target ex:EHR ;
    odrl:assigner ex:provider ;
    odrl:assignee ex:requester ;
    odrl:permission [
        odrl:constraint [
            odrl:and ex:constraint_purpose, ex:constraint_temporal ] ;
        odrl:duty [ odrl:action duodrl:CollaborateWithStudyPI ] ] .

ex:constraint_purpose a odrl:Constraint ;
    odrl:leftOperand oac:Purpose ;
    odrl:operator odrl:isA ;
    odrl:rightOperand duodrl:HMB .

ex:constraint_temporal a odrl:Constraint ;
    odrl:leftOperand odrl:elapsedTime ;
    odrl:operator odrl:lteq ;
    odrl:rightOperand "2024-12-31"^^xsd:date .
```

matching systems are supposed to match offers with requests using subsumption as a criterion, the Algorithm 6.1, proposed for OAC-based systems in the previous Section, can also be adapted here for the specific use case of DUO-based health data-sharing. In DUO's case, for a class P and its subclass C, a request to access P does not allow access to C since C is more specific than P. However, a request to access C would allow use of P as P is less specific than C. In light of these considerations, Algorithm 6.2 offers pseudo-code that showcases the adapted steps of the policy matching algorithm for the DUO health data use case.

In this context, it is important to note that this algorithm offers a general outline of actions, since DUO's documentation lacks specific details for accurately interpreting certain semantic aspects. Since this interpretation significantly influences decision-making within DUO's processes, and DUO only specifies how to interpret purposes and not other restrictions, e.g., location, users, projects, this is as an area that might require further exploration and investigation in the future as more details about DUO's matching process are known. To address this information gap, apart from the purposes in DUO's concepts, the matching of the remaining concepts is performed according to the proposed Algorithm 6.1, as this proposal considers a legally relevant interpretation of hierarchical concepts, where a narrower concept cannot be deemed compatible with a request for a broader concept, e.g., a permission to access the data within a certain city cannot be fulfilled by a request for a region broader than that city. Hence, the proposed algorithm mirrors the previously described algorithm for OAC. Firstly, the prohibitive statements are matched, and only if no incompatibilities are found are the permissions considered. The denial of the access request occurs during prohibition checking if any of the following restrictions in the dataset offer are found to be

Algorithm 6.2 Pseudo-code of the proposed matching algorithm for DUODRL.

for $prohibition \leftarrow odrl{:}Offer$ **do**
  if $odrl{:}assignee \in offer{:}prohibition$ **then**
    if $offer{:}assignee \equiv request{:}assignee$ **then** $decision \leftarrow DENY$
  for $constraint \leftarrow prohibition$ **do**
    if $odr{:}spatial \leftarrow constraint$ **then**
      if $offer{:}spatial \cap request{:}spatial \neq \varnothing$ **then** $decision \leftarrow DENY$
    else if $duodrl{:}Project \leftarrow constraint$ **then**
      if $request{:}project \cap offer{:}project \neq \varnothing$ **then** $decision \leftarrow DENY$
    else if $odrl{:}dateTime \leftarrow constraint$ **then**
      if $timeNow < moratoriumDate$ **then** $decision \leftarrow DENY$
    else if $offer{:}purpose \cap request{:}purpose \neq \varnothing$ **then** $decision \leftarrow DENY$
for $permission \leftarrow odrl{:}Offer$ **do**
  if $odrl{:}assignee \in offer{:}permission$ **then**
    if $offer{:}assignee \nequiv request{:}assignee$ **then** $decision \leftarrow DENY$
  for $constraint \leftarrow permission$ **do**
    if $odrl{:}dateTime \leftarrow constraint$ **then**
      if $timeNow > timeLimit$ **then** $decision \leftarrow DENY$
    else if $request{:}purpose \in groupResearchPurposes$ **then**
      if $request{:}purpose \notin offer{:}purpose \bigvee request{:}group \notin offer{:}group$ **then** $decision \leftarrow DENY$
    else if $request{:}purpose \notin offer{:}purpose$ **then** $decision \leftarrow DENY$
if $\nexists DENY$ **then** $decision \leftarrow GRANT$

incompatible with the data request:

1. offer assignee matches (≡) the request assignee;
2. offer has a spatial constraint matching or not satisfying ($\cap \neq \varnothing$) the spatial constraint in the request;
3. offer has a project constraint matching ($\cap \neq \varnothing$) the project in the request;
4. there is a moratorium period with a date in the future; and
5. offer has a purpose matching ($\cap \neq \varnothing$) the purpose in the request.

If no prohibitions are identified, the next step is to verify the permissions. Access will be denied during permission checking if any of the following constraints in the request are incompatible with the dataset offer:

1. offer assignee does not match (≢) the assignee of the request;
2. offer time limit on use has lapsed;
3. offer has a group-related research purpose, e.g., `PopulationGroupResearch`, `AgeCategoryResearch` or `GenderCategoryResearch` and the request purpose does not match (∉) it or the request purpose matches it but the group does not (∉), e.g.

the `PopulationGroup`, `Age` or `Gender` in the request are different from the one in the offer; and

4. offer purpose does not match ($\notin$) request purpose, e.g., DUO's general research use purpose GRU in a request does not match a health, medical or biomedical research purpose HMB in an offer as GRU is a superclass of HMB.

The described procedures are applied to all permissions and prohibitions outlined in the dataset's offer. If all permissions and prohibitions are met without any violations, access to the health-related data can be authorised, if not the `odrl:Agreement` will include a prohibition that states the motives for the denial of the request. The proof-of-concept implementation outlined in Section 6.3.3 applies these steps to match an offer with request policies in order to generate a data access agreement.

Table 6.3 illustrates examples of how the matching algorithm operates for prohibitions and permissions in offers with spatial and purpose-based constraints. As previously mentioned, in a semantic-based context, performing equivalence ($\equiv$), intersection ($\cap$), and subset ($\subseteq$) checking requires supplementary deliberations beyond simply employing `owl:sameAs` or `rdfs:subClassOf` inferences. For instance, comparing an offer with a location constraint restricted to Spain with a request to Europe using the subset ($\subseteq$) operator for permissions or the intersection ($\cap$) operator for prohibitions needs both locations to be articulated in a manner that facilitates such a hierarchical interpretation. In this scenario, the matching process requires understanding that *Spain* is a narrower concept, geographically speaking, than *Europe*. Further complexities arise when representing legal jurisdictions, such as the *EU*, of which *Spain* is a member.

Table 6.3: Examples of outcomes of the policy matching algorithm for DUODRL.

Offer			**Request**		**Outcome**	
Rule	Purpose	Location	Purpose	Location	Decision	Reason
Permission	GS	Spain	GS	Europe	DENY	Europe $\notin$ Spain
Permission	GS	Europe	GS	Spain	GRANT	Spain $\subseteq$ Europe
Prohibition	GS	Spain	GS	Europe	DENY	Europe $\cap$ Spain $\neq \varnothing$
Prohibition	GS	Europe	GS	Spain	DENY	Spain $\cap$ Europe $\neq \varnothing$
Prohibition	GS	UK	GS	Spain	GRANT	Spain $\cap$ UK $= \varnothing$
Permission	HMB		DS-Cancer		GRANT	DS-Cancer $\subseteq$ HMB
Prohibition	DS-Cancer		HMB		DENY	HMB $\cap$ DS-Cancer $\neq \varnothing$

Expressing legal compliance with OAC The terminologies and concepts employed in DUO differ significantly from those utilised in legal compliance tasks, e.g., such as the ones mandated by the GDPR. With the proposed DUODRL approach which leverages ODRL concepts, the terminologies involved are specified in a language which can be legally interpreted, e.g., such as the `Asset` or `Party` terms. Moreover, the ODRL vocabulary encompasses complementary terms that could extend DUO with precise legal annotations, such as the `odrl:consentingParty`, `odrl:informedParty`, and `odrl:obtainConsent` terms. While these terms suffice for endowing a legally-aligned policy, they may not adequately encompass the specific requirements of laws such as the GDPR, which assigns distinct roles to entities and mandates the use of specific

legal grounds for the processing of personal data, including additional requirements for special categories of such data, e.g., such as health data. Simultaneously, tailoring the terms solely for a single jurisdiction could confine the usefulness and applicability of the proposed work solely to that law, without providing a clear pathway for accommodating other laws and jurisdictions. To bridge this divide, this Thesis proposes the usage of the OAC profile, which provides legally-aligned terms designed to be jurisdiction-agnostic and adaptable across various legal frameworks.

The work on OAC can be seamlessly integrated into this use case as it already aligns ODRL terms with DPV terms for the expression of legally-aligned policies. Using OAC, and by consequence DPV, enables the modelling of policies with constraints on the legal basis used to process the data, e.g., consent, or on the third party recipients of the results of the processing, as well as explicitly define the entities processing the data – the data controllers – and the data protection laws that must be respected. The `oac:TechnicalOrganisationalMeasure` constraint left operand can also be used to ensure that certain measures are taken before or after data access, e.g., impact assessments or records of activities. To explicitly designate GDPR as the law applicable to the policy and rely on its legal bases for processing, DPV's GDPR extension can be used. Through this segregation (between DPV and its GDPR extension), policies can be articulated in a jurisdiction-agnostic way, and subsequently tailored to a specific law such as GDPR by factoring in additional contextual information, such as the locations of patients whose data is implicated or the identity of the legal representative the requesting data controller. This separation also paves the way for applying other laws, within or outside of the EU, which would be possible by crafting other law extensions to DPV akin to developed GDPR one. Listing 6.6 provides two ODRL offer policies: the first uses DPV to enact jurisdiction-agnostic data protection terms and the second uses the GDPR extension to describe GDPR-specific terms.

In its documentation, DUO specifies that it is the responsibility of the DUO adopter to interpret and apply data protection requirements such as the mandated GDPR rights and obligations. This acknowledgement stems from the challenges associated with determining their relevance prior to receiving a data request or due to variations in provider/requester jurisdictions. To aid in this endeavour, this Thesis proposes incorporating or providing relevant methodologies essential for identifying GDPR's applicability (or that of other laws). For instance, GDPR is deemed applicable when an organisation does its business within the EU or handles the personal data of EU individual. This entails having the knowledge of the data subjects' birthplace of individuals whose data is being offered for use, as well as the location of the data controller.

Both scenarios, jurisdiction-agnostic and GDPR-specific, can be implemented into ODRL policies through the relevant `dpv:Entity` and `dpv:Location` concepts. This facilitates the expression of further data access conditions using ODRL, such as data accessibility depending on the data request acknowledging GDPR's applicability or allowing access solely within GDPR-regulated jurisdictions. These conditions can be encoded as permissions or prohibitions to be checked during the matching step, similarly to other constraint checks described in the algorithm outlined in this Section. DPV's legal extension, which provides data protection laws and authorities taxonomies, should be used to specify such conditions. Moreover, the possibility of having a jurisdiction-agnostic scenario, while still maintaining a strong connection with general legal requirements, alleviates the burden on adopters who may prefer not to express such information with jurisdiction-specific details. For example, a data provider who simply mandates data access

Listing 6.6 `odrl:Offers` that use DPV and its GDPR extension to indicate conditions of access to data. `ex:offer_agnostic` is jurisdiction-agnostic and requires consent and an impact assessment. `ex:offer_gdpr` is GDPR-specific and requires explicit consent and a DPIA.

```
ex:offer_agnostic a odrl:Offer ;
    odrl:uid ex:offer_agnostic ;
    odrl:profile oac: ;
    rdfs:label "Offer to use dataset for GRU using consent, and requiring
     ↪ an impact assessment" ;
    dcterms:source duodrl:DUO_0000042 ;
    dcterms:issued "2024-04-26"^^xsd:date ;
    odrl:target ex:EHR ;
    odrl:action oac:Use ;
    odrl:assigner ex:provider ;
    odrl:assignee ex:requester ;
    odrl:permission [
        odrl:constraint [
            odrl:leftOperand oac:LegalBasis ;
            odrl:operator odrl:isA ;
            odrl:rightOperand dpv:Consent ] ] ;
    odrl:permission [
        odrl:constraint [
            odrl:leftOperand oac:TechnicalOrganisationalMeasure ;
            odrl:operator odrl:isA ;
            odrl:rightOperand dpv:ImpactAssessment ] ] .
ex:offer_gdpr a odrl:Offer ;
    odrl:uid ex:offer_gdpr ;
    odrl:profile oac: ;
    rdfs:label "Offer to use dataset for GRU using GDPR's explicit
     ↪ consent, and requiring a DPIA" ;
    dcterms:source duodrl:DUO_0000042 ;
    dcterms:issued "2024-04-30"^^xsd:date ;
    odrl:target ex:EHR ;
    odrl:action oac:Use ;
    odrl:assigner ex:provider ;
    odrl:assignee ex:requester ;
    dpv:hasDataSubject ex:provider ;
    dpv:hasDataController ex:requester ;
    dpv:hasApplicableLaw legal-eu:law-GDPR ;
    odrl:permission [
        odrl:constraint [
            odrl:leftOperand oac:LegalBasis ;
            odrl:operator odrl:isA ;
            odrl:rightOperand eu-gdpr:A6-1-a-explicit-consent ] ] ;
    odrl:permission [
        odrl:constraint [
            odrl:leftOperand oac:TechnicalOrganisationalMeasure ;
            odrl:operator odrl:isA ;
            odrl:rightOperand dpv:DPIA ] ] .
```

based on consent, without explicitly defining the conditions for valid consent under the GDPR, can articulate this policy using DUODRL and the previously described OAC profile, while the data access ethics board can assess it further based on their understanding of valid consent requirements. As such, this organisation can remove the burden of GDPR requirements from the data subject and impose any necessary additional restrictions or obligations directly to the data requester, to ensure compliance with this data protection regulation, e.g., using DPV's GDPR extension.

This flexibility also provides advantages for systems like DUOS, which can explicitly designate datasets as necessitating GDPR-level consent or indicating its applicability by incorporating relevant metadata into the dataset offer. This approach aids the matching process in verifying legal obligations and compatibility. For instance, it may require specific information about the requester, e.g., such as information regarding the identity of their DPO, or demand additional legal bases and safeguards for data transfer to outside of the EU. Through this approach, DUO and its users and adopters can extend their legal applicability globally and also possess the means and mechanisms to address specific requirements of specific laws.

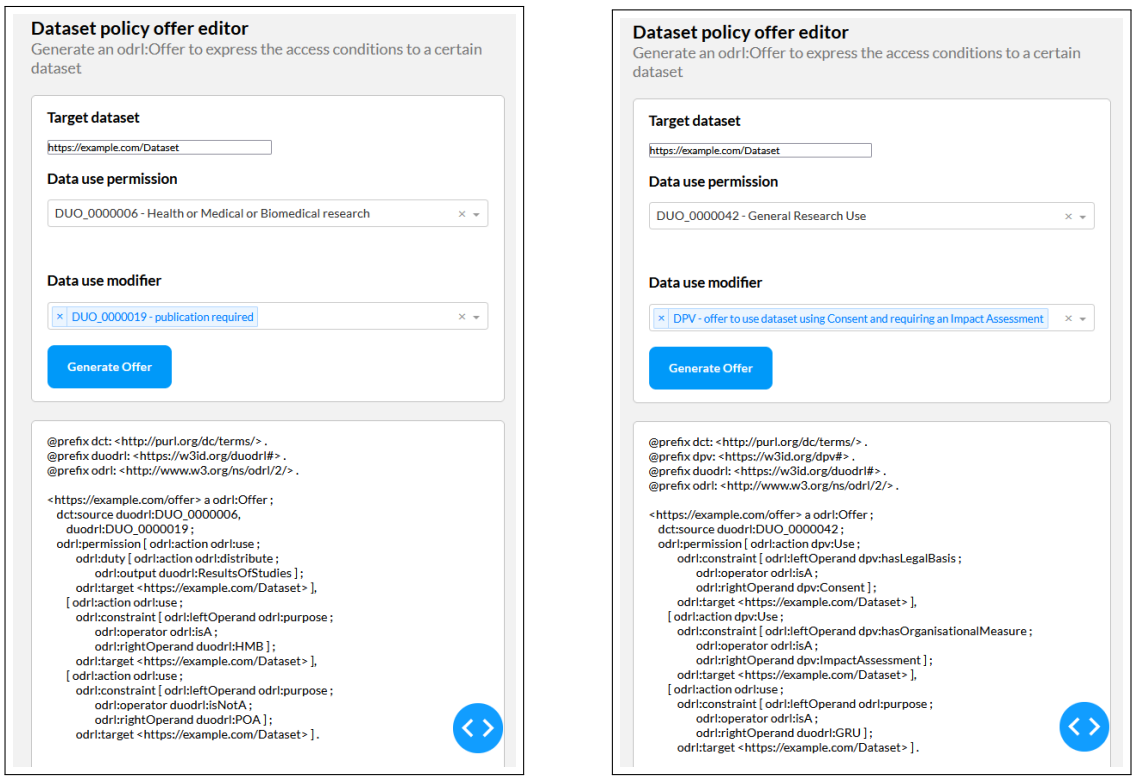
6.3.3 Proof of Concept implementation

In this Section, the implementation of a Proof of Concept designed for generating DUODRL policies is outlined. This PoC also includes the implementation of the policy matching algorithm described in the previous Section, which can be accessed at `https://w3id.org/duodrl/demo`. Moreover, the PoC is installable and usable locally, while an online demonstration is also accessible at `https://w3id.org/duodrl/app`.

Figure 6.8 depicts two instances of the developed UI of the PoC, designed for editing DUODRL policies. Both examples in the Figure use ODRL and the developed DUODRL extension to generate DUO-based policies for access to health datasets. Example (a) exclusively employs DUODRL concepts, while example (b) integrates both OAC and DPV to formulate legally-aligned `odrl:Offer`s. When selecting the relevant data use permission and modifiers in the UI, the PoC application fetches the corresponding `odrl:Set`s and instantiates them as an `odrl:Offer` policy associated with a particular dataset. This policy is presented to the user in the UI as well. The source code of the appliaction and the DUODRL policies utilised for this purpose are accessible online at `https://w3id.org/duodrl/repo`. Detailed instructions on how to install, launch, and examples of usage of the PoC application are also available on the source code repository at `https://w3id.org/duodrl/demo`.

As mentioned in the last Section, the matching algorithm employed in this PoC was adapted from the OAC-based algorithm presented in Section 6.2.3 to cater to DUO's use case health data-sharing conditions, while ensuring alignment with legal requirements. During the matching process, the conditions outlined in the request for health data, in the `odrl:Request` policy, must be compared with those specified in the health datasets' offers, in the `odrl:Offer` policy. This entails ensuring that the permissions and prohibitions from the offer instantiation can be met by the request. Once this compatibility is established, the policies are deemed congruent, a permissive `odrl:Agreement` is generated, and access can be granted according to the established conditions. If the policies are not compatible then a prohibitive `odrl:Agreement` is generated and access to the dataset is denied. Moreover, for data discovery purposes, the

Figure 6.8: Proof of concept implementation showing generation of `odrl:Offer` policies



(a) with DUODRL concepts.

(b) with OAC and DPV concepts.

`odrl:Request` policy needs to be compared against the policies of each available dataset if this PoC is to be implemented within a large-scale system, to ensure that all available datasets and corresponding policies are checked to see if they fit the request. While pre-computations and optimisations could streamline this process, this is not implemented within this Thesis prototype, as its purpose is solely to serve as a first PoC to test the generation of machine and human-readable data access agreements.

The policy matching algorithm initiates by verifying if a dataset's policy includes a specific rule corresponding to the purpose outlined in the `odrl:Request`. If the dataset's offer has a prohibition for a purpose P, access to the dataset may be denied if the request is for P; conversely, if a permission is detected, access can be allowed allowed. Subsequently, as described in the previous Section, similar assessments should be conducted to examine other restrictions beyond purpose, i.e., the ones outlined in Table 6.2, such as constraints on the type of assigner for the agreement, on the location or timing of data use. In the event of encountering a matching prohibition, access to the dataset must be refused; however, if a permission is present, access may be granted.

If additional obligations are mandated for dataset access and use, such as committing to collaborate with the primary study investigator or providing proof of ethical approval to perform the study, these should included within the `odrl:Agreement` together with the conditions for access. In instances where conflicting policies arise from the merge of distinct data use permissions and modifiers, the prohibition supersedes by default, akin to the algorithm's base protocol, established

in Section 6.2. Under such circumstances, access to the dataset is declined.

The outcome of the matching algorithm, an `odrl:Agreement`, should be generated, incorporating a rule either permitting or prohibiting the solicited `odrl:Request`, as well as other metadata described in Section 6.3.2, including linkage with the corresponding `odrl:Request` and `odrl:Offer`, and any duties, if specified by the data provider. In this way, the `odrl:Agreement` instantiation serves to encapsulate and facilitate the creation of appropriate legal documentation to formalise and communicate the agreement stemming from the submitted request and consequent matching outcome. The computational overhead associated with generating such an agreement, particularly at a time when the request will be matched against a large number of offer policies, falls outside the scope of analysis of this Thesis. This is due to the fact that the overhead will vary significantly depending on the use case and on the interpretation of the DUO concepts, which are not explicitly defined and, as such, reflect the specific analysis proposed in this Thesis, aligned with its core foundation on the decentralisation of access to data. As a result, the focus of this Section is on clearly representing the information through the utilisation of ODRL policies and showcasing the applicability of the proposed OAC-based algorithm for decentralised access to health data on a purpose-based setting, aligned with the legal requirements specified in the GDPR.

PoC preservation, maintenance, and future improvements The PoC implementation is published and archived according to the methodology described in Section 3.7.3. Furthermore, its source code is hosted at `https://w3id.org/duodrl/repo`, under the CC-BY-4.0 license. A live demonstration of the PoC UI features is also available at `https://w3id.org/duodrl/app`. The repository can also be used by DUO/DUODRL users to suggest new features to be added to the PoC, as well as to report bugs through GitHub Issues.

6.4 Evaluation and concluding remarks

In this Section, the developed work is evaluated by assessing the quality of the implementation of DUODRL, using OOPS! to detect common issues with ontologies and FOOPS! to check the alignment with FAIR principles, and by comparing the proposed policy-based algorithm with DUO's and Solid's existing access control systems. Concluding remarks concerning the architecture, algorithms, and PoC implementation proposed in this Thesis are also specified at the end of the Chapter.

6.4.1 Ontology quality evaluation

This Section describes the outcomes of DUODRL's quality evaluation, including the detection of common pitfalls with OOPS! [Poveda-Villalón et al., 2014] and the alignment with FAIR principles with FOOPS! [Garijo et al., 2021].

In terms of quality evaluation, the OOPS! tool was used to detect common errors in ontology development, such as missing domain or range properties or missing human-readable annotations. Through the OOPS! evaluation no critical nor important issues were detected. Furthermore, FOOPS! was used to evaluate the alignment of the developed vocabularies with the FAIR principles. The following results were obtained:

- Findable – 8/9
- Accessible – 2/3
- Interoperable – 2/3
- Reusable – 8.83/9
- FOOPS! overall score – 91%

These outcomes are aligned with the scores obtained for OAC and PLASMA, and largely exceed the ones computed for DUO (Findable – 4.50/9; Accessible – 2/3; Interoperable – 3/3; Reusable – 3.50/9; FOOPS! overall score – 54%). Furthermore, DUODRL obtained a good score in all FAIR aspects. In terms of improvements, DUODRL can be submitted to LOV for it to be recorded in a public registry of ontologies. This will improve both the findability and the accessibility of the vocabulary.

6.4.2 Comparison with existing access control systems

In this Section, the proposed policy matching algorithm is compared with DUO's system, as this is being used in production by organisations to share health-related datasets for research purposes, and with Solid's current access control mechanisms, as this was the environment chosen in this Thesis to demonstrate the utility of the proposed policy-based access control system due to its decentralised and Semantic Web based nature. Table 6.4 presents the results of this comparison.

The different systems are compared according to their ability to represent different types of policies, e.g., requirement, preference, offer, request, and agreement, different types of rules, e.g., permissions, prohibitions, and duties, and different types of restrictions and constraints, from legal requirements related to the type of data and processing operations to use case specific restrictions such as spatial, gender or age constraints. As it can be checked in Table 6.4, Solid's WAC system provides the lowest level of complexity in terms of access control, as it only allows the expression of authorisation statements, which can be interpreted to signify both the user's offers as well as the agreements to access data. Moreover, it only allows the expression of permissive statements, the processing operations are restricted to access actions, e.g., read, write, append and control, and available constraints are related to assignees and applications requesting access to data. Solid's ACP improves this system by allowing the expression of requests for data, through context graphs, as well as the definition of prohibitive statements and constraint on the identity provider of the data requester.

Using this analysis, DUO's matching system can be considered more advanced than the current Solid efforts on access control, as beyond covering permissive and prohibitive statements, it also includes concepts to express duties on the data requesters. A feature which both DUO and Solid's access control mechanisms share is related to the fact that access can only be allowed and/or denied to resources/datasets and not to particular types of data – this is flagged in Table 6.4 using the $*$ character. DUO's concepts also allow the matching of constraints related to the assignee of the data request, purpose, temporal and spatial restrictions, as well as use case-related restrictions such as the project in which the data is going to be used, the gender, age group or population group of the providers of the data, and, in case where the purpose of the request is to perform

Table 6.4: Comparison of proposed in this Thesis with other access control systems.

	Algorithm approaches			
Feature	**Thesis**	**DUO**	**ACP**	**WAC**
Requirement	OAC	Data Use Permission, Data Use Modifier		
Preference	OAC			
Offer	ODRL		ACR	Authorisation
Request	ODRL	Investigation	Context	
Agreement	ODRL		Access grant	Authorisation
Permissions	ODRL	Data Use Permission	allow	✓
Prohibitions	ODRL	Data Use Modifier	deny	
Duties	ODRL	Data Use Modifier		
Data	OAC/DPV	*	*	*
Processing	OAC/DPV		Read, Write, Append, Control	Read, Write, Append, Control
Assigner	ODRL			
Assignee	ODRL	✓	agent	agent, agent group
Application	OAC		client	origin
Service	OAC			
Legal roles	OAC/DPV			
Purpose	OAC/DPV	✓		
Legal basis	OAC/DPV			
TOM	OAC/DPV	✓		
Technology	OAC/DPV			
Identity provider	OAC		issuer	
Spatial	ODRL	✓		
Project	DUODRL	✓		
Temporal	ODRL	✓		
Disease	DUODRL	✓		
Gender	DUODRL	✓		
Age	DUODRL	✓		
Population	DUODRL	✓		

research on specific diseases, the disease being studied. These features of the DUO policy matching process are flagged with a ✓ since the specific details of DUO's implementation of the algorithm are unknown.

To conclude, even though DUO's access control mechanism provides ways to express most of the features identified in Table 6.4, the approach proposed in this Thesis supersedes it as the Thesis' algorithms are not just focused on health data sharing, e.g., can be used to control access to other types of personal data for purposes beyond research, and allow for the expression of the highest number of different types of policies and constraints on policy rules, including legally-aligned constraints, e.g., the legal basis used to justify the processing of personal data. The Thesis proposed solutions to tackle each feature of the policy matching algorithm, i.e., ODRL, DPV, OAC, or DUODRL, are also included in Table 6.4 to showcase how the developed work contributes to different aspects of the policy algorithm design.

6.4.3 Concluding remarks

The decentralised nature of Solid provides a good ground for improvements of its access control mechanism towards having a legally-aligned ecosystem. By adding fine-grained policies, as well as other provenance metadata to the architecture of such environments, transparency and accountability are improved for both data subjects and data controllers. Moreover, the addition of a policy matching algorithm paves the way for future, real-time negotiation of access and usage conditions of data, e.g., through Web agents, which on one hand will aid data subjects on actively controlling only what they want to control, while data controllers have a mechanism and audit trail to justify how they got access to what data under which conditions. As future work, supported by the under-development ODRL formal semantics specification, explainability features should be integrated into the algorithm so that data subjects can understand, through a human-readable interface, how it took certain decisions, e.g. why the prohibitions were prioritised over the permissions. Other PoCs should be performed for other types of data, as well as focusing on other data sharing scenarios beyond data sharing for research purposes. Furthermore, the evolution of this work, towards production-ready architectures for the decentralised access to personal data, should continue to be followed by both technical and legal experts to ensure that any changes on the technical or legal level, e.g., changes made to policy languages or vocabularies, or new laws being proposed, respectively, can be easily accommodated.

Chapter 7

Going beyond the GDPR – Exploring the Data Governance Act

The content of this Chapter has already been partially included in the articles published during this Thesis [Esteves and Rodríguez-Doncel, 2022b, Esteves et al., 2023, Esteves, 2023].

The source code produced during the development of this chapter is stored at:

- https://w3id.org/dgaterms/repo
- https://w3id.org/people/besteves/soda/repo

After witnessing the influence of the GDPR on managing the personal data of European citizens, the European Commission has shifted its attention to establishing a unified data strategy. This strategy aims to encourage the (re)use and exchange of data among citizens, businesses, and governments, all while ensuring control remains with the original data generators [European Commission, 2020]. In this context, a new series of regulations concerning the usage and management of data is currently under consideration for adoption across the EU countries. Within this framework, the Data Governance Act (DGA) [2022g], a regulation primarily focused on regulating the operations of data intermediation services and data altruism organisations, was approved by the Commission in May 2022 and is now mandated for implementation in all EU member states.

This Chapter outlines a range of requirements outlined by the DGA aimed at safeguarding the interests of both data subjects and data holders – a new legal role inserted in the DGA to refer to entities providing non-personal data –, as well as regulating the jurisdiction of relevant authorities. Additionally, it presents various scenarios where the application of Semantic Web technologies, such as the developed work on OAC, ODRL, and DPV, could assist these stakeholders in meeting their respective new rights and responsibilities. More specifically, it aims to achieve three main goals: (i) generating machine-readable policies for the reuse of public data, (ii) defining consent and permission terms for data altruism, and (iii) establishing standardised registers of data altruism organisations and intermediation service providers and records of their activities. By leveraging these semantic vocabularies, the aim is not only to enhance machine-readability and interoperability but also to streamline the modelling of data-sharing policies and consent forms across various

scenarios, as well as facilitating the creation of a shared semantic framework for maintaining public registers of data intermediaries and altruism organisations, along with documenting their activities. Given the accessibility and extensibility of these vocabularies, adapting them to meet specific requirements outlined in the DGA becomes a straightforward task. The prefixes and namespaces used in the Listings in this Chapter are explicitly defined in the Namespaces list.

Section 7.1 provides a description of the DGA and the entities within it, as well as the identification of flows of information between entities. Moreover, three scenarios where Semantic Web technologies can be leveraged to assist data subjects and data holders in establishing the conditions for the reuse of their data, and public sector bodies, data intermediaries, and altruistic organisations in fulfilling their DGA-related obligations.

Section 7.2 discusses the usage of ODRL, DPV, DCAT and other vocabularies to express policies for the reuse and sharing of public sector body-held data, to represent registers of entities and logs of their activities, and to create uniform data altruism forms to record consent actions of data subjects and permissions of data holders. Based on the identified gaps, Section 7.3 describes the development and evaluation of DGAterms, a vocabulary for the representation of DGA requirements.

Section 7.4 describes the development of a Solid Data Altruism application, to implement data altruism as a service using Solid and ODRL policies to grant access to personal data for altruistic purposes in a privacy-friendly manner.

To conclude, Section 7.5 debates the advantages and challenges of the proposed semantic model towards having semantic interoperability to support the development of common European data spaces.

7.1 Information flows in the DGA

In this Section, a detailed description of the DGA and its core legislative objectives is provided, including the introduction of use cases where Semantic Web technologies and decentralised data environments can be leveraged to aid data intermediaries and data altruism organisations to implement their services while fulfilling their renewed legal duties, in particular, related to the reporting of their activities and how they use personal and non-personal data from data subjects and data holders, respectively.

In February 2020, the European Commission [2020] introduced a series of regulatory proposals aimed at legislating the European strategy for data, encompassing a set of new regulatory proposals aimed at governing the utilisation of non-personal and public data, regulating digital services and markets, and fostering the creation of common European data spaces. By prioritising data and ensuring its accessibility across sectors, this transformation must also maintain the interests of both data subjects and data holders, while supporting trusted entities to facilitate data sharing aligned with new regulations. Among these proposals was the Data Governance Act [2022g], a regulation presented to enhance data accessibility and foster trust in data intermediation services throughout the EU. Following approval by both the European Parliament and the European Council, this legislation entered into force on 23 June 2022 and, following a 15 month grace period, has been applicable since September 2023. Similarly to other data-related legislation within the EU, the DGA relays new rights and obligations to entities holding both personal and non-personal data. It

also regulates the operations of data users and two categories of data-related services related to data intermediation and altruism. The principal objectives of this legislation include:

(i) Facilitating the reuse of protected public-sector data while maintaining its privacy and confidentiality, particularly in cases where such data is subject to the rights of others, including trade secrets, personal data protection, and data safeguarded by intellectual property rights.

(ii) Regulating and maintaining a register of data intermediation service providers, which facilitate data sharing among enterprises and support individuals to have a 'personal data-sharing intermediary', designed to aid them in exercising their rights, e.g., under the GDPR.

(iii) Allowing businesses and data subjects to voluntarily contribute data for altruistic purposes, such as medical research.

(iv) Establishing a novel supervisory authority, the European Data Innovation Board (EDIB), tasked with supervising the operations of data intermediation service providers and data altruism organisations. This Board comprises representatives from regulatory bodies overseeing data intermediation and altruism activities across all EU member states, from the EDPB and EDPS, from the European Union Agency for Cybersecurity (ENISA) and the European Commission, as well as experts in standardisation, portability, interoperability, and other pertinent stakeholders.

The main hurdles to overcome in order to achieve these objectives are associated with:

(i) *Availability/Discovery of datasets*: in the absence of technical assistance for creating data spaces and reliable data-sharing platforms, individuals and organisations will not have tools to share their data for the common good, nor will they have adequate support to exercise their data-related rights. On the other hand, data users lack the necessary tools to search for the data they require.

(ii) *Establishment of data access and usage conditions*: with the unavailability of standards and metadata vocabularies to articulate machine-readable policies, setting conditions for the usage and access to personal, non-personal, and public-sector data – rooted not just in legal frameworks but also in ethical, organisational, and social norms – will lead to interoperability challenges among entities providing and seeking data access.

(iii) *Reporting duties*: without maintaining structured records of their activities, providers of data intermediation services and organisations engaged in data altruism will depend on manual methodologies to generate documentation reporting their accountable and responsible data-handling practices.

Thus, to tackle the challenges at hand, the first task defined in this Thesis is related to the identification of relevant flows of information between DGA-regulated entities, as well as what specific information items need to be shared or kept by which entities. Since the DGA both advocates for data availability and legislates data sharing, a delineation of information flows among data-sharing entities can be specified. In addition, as a result of these interactions, certain registers and records of activities must be maintained for transparency and accountability, in line with previous EU data protection law, e.g., the GDPR. Figure 7.1 outlines the entities and their corresponding information flows identified within this context. Their respective definitions are

outlined below.

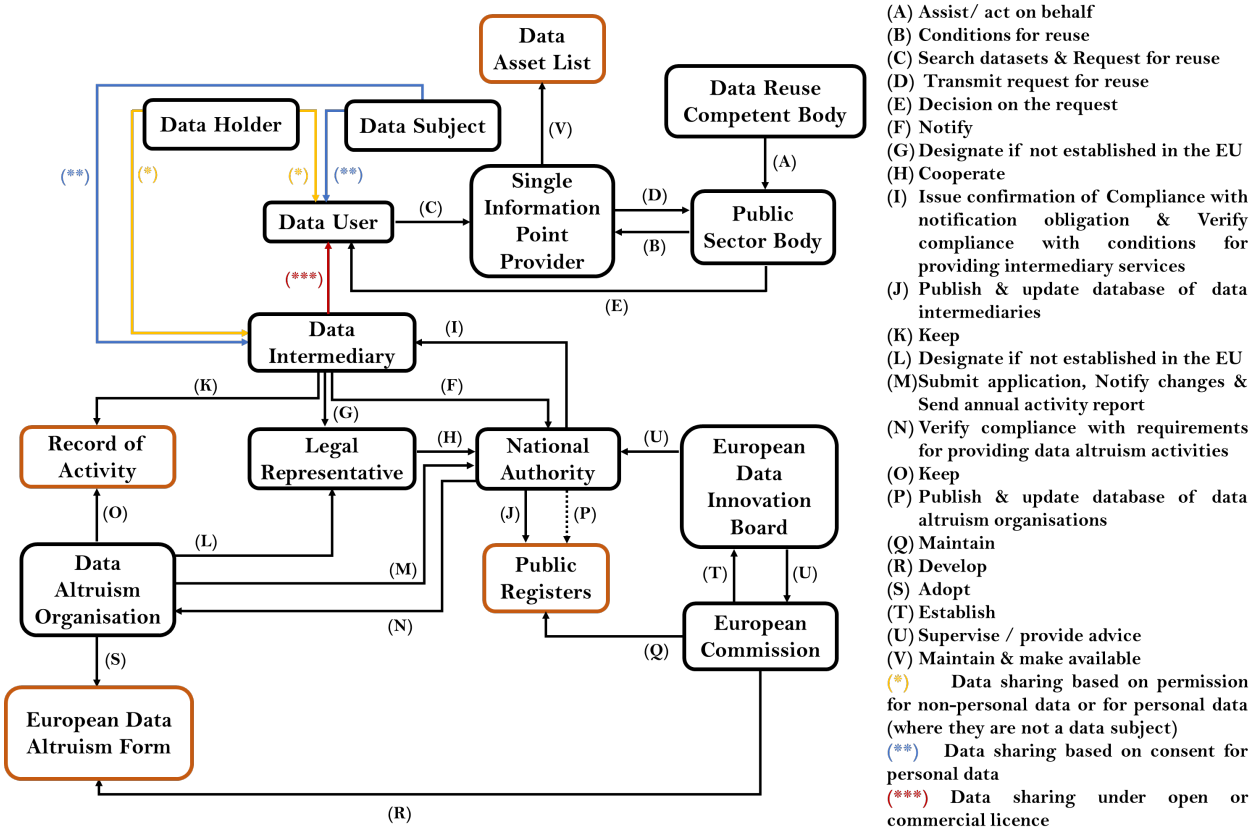


Figure 7.1: Flows of information between DGA entities, adapted from Esteves et al. [2023]. The concepts in a black box are the entities and the ones in an orange box are the legal documentation that needs to be created and maintained by said entities. The direction of the arrows represents the direction of the information flow between entities. A short description of each flow is specified on the right side of the Figure.

Data Subject Natural person whose personal data is undergoing any kind of processing (as defined in the GDPR).

Data Holder An entity that can share personal and/or non-personal data.

Data User An entity that wishes to use personal and/or non-personal data for commercial or non-commercial purposes.

Data Intermediation Service Provider An entity that establishes commercial relationships for the sharing of data among data subjects, data holders, and data users.

Data Altruism Organisation An not-for-profit organisation that collects and shares data for altruistic purposes.

Public Sector Body An entity or group of entities governed by public law from one or more State, regional and/or local authorities.

Legal Representative A legal entity's representative appointed to act on behalf of a data intermediation service provider or altruistic organisation.

Competent Body An entity appointed by a public sector body to offer legal and technical assistance regarding the access and reuse of public sector data.

Single Information Point Provider An entity tasked with receiving and forwarding requests for public data reuse.

Competent Authorities Authorities in charge of supervising the activity of data intermediation service providers and data altruism organisations and maintaining a public register of said entities.

European Data Innovation Board A supranational authority responsible for supervising the operations of data intermediaries and data altruism organisations.

As illustrated in Figure 7.1 by a red arrow, the data intermediation service provider, or data intermediary, provides data users with access to data under open or commercial license conditions and, as illustrated with the (K) arrow, must maintain a record of its activities. This diagram was crafted from an analysis of DGA's Chapters II ('Re-use of certain categories of protected data held by public sector bodies'), III ('Requirements applicable to data intermediation services'), IV ('Data altruism'), and VI ('European Data Innovation Board') [2022g]. Each article within these chapters was meticulously examined to identify interactions among the identified entities. Whenever an information flow was identified between multiple entities, the respective interaction was documented in the diagram. Furthermore, requirements related to compliance documentation are also included in the diagram. These requirements entail the recording of information that can be automated using semantic technologies.

In the following three Sections, the information flows concerning the conditions for reusing public data (Section 7.1.1), the maintenance of registers of altruistic and intermediary entities and respective activities logs (Section 7.1.2), and the forms to record data altruism terms (Section 7.1.3) are described in detail. These areas highlight the strengths of using semantic technologies to effectively aid the previously mentioned entities in automating their DGA compliance tasks. For each example use case, a systematic analysis of the pertinent information flows and the corresponding data exchange requirements was conducted manually. The findings are organised and presented in the subsequent Sections.

7.1.1 Conditions for the reuse of data held by public sector bodies

DGA's Chapter II legislates the reuse of data stored by public sector bodies, encompassing the specification of which data categories are regulated (Article 3 [2022g]), the conditions required from public sector bodies to provide such services (Articles 5 and 6 [2022g]), and the delineation of single information point providers and their role in facilitating data users' search for and request of datasets for reuse purposes (Articles 8 and 9 [2022g]). To keep such data, these entities need to have in place safeguards to protect their commercial and statistical confidentiality, intellectual property rights of third parties, and personal data-related rights. Moreover, they have to publish the dataset reuse conditions, and the respective data request procedure, in a transparent and publicly accessible manner. The reuse must be contracted, with a maximum duration of one year, including the categories of data being used as well as the purpose for the reuse. Public sector bodies

also have the right to obtain guidance and technical support from a competent body, which must be appointed by each EU member state. These appointed entities are responsible for providing guidance on data formats and storage, assisting in the implementation of privacy-preserving methods to protect personal data integrity, and supporting activities to obtain consent from data subjects and permission from data holders.

A complete list of information required from public sector bodies, along with a reference to the pertinent articles in the DGA, is summarised in Table 7.1. This information, which may be further specified with the aid of a competent body (as previously mentioned and indicated by the (A) arrow in Figure 7.1), should be communicated to the single information point provider (as depicted by the (B) arrow). This enables data users to search for datasets ((C) arrow) and submit reuse requests through the single information point ((D) arrow). Additionally, single information point providers are obligated to maintain and disclose a data asset list (illustrated by the (V) arrow), comprising details on available resources and their reuse conditions.

Table 7.1: Information items about public sector bodies' services.

Article	Information items
2.9	Data user/categories of users
5.1	Public sector body information
5.1	Competent body information
5.2	Categories of data
5.2	Purposes for usage and access
5.2, 5.3(a)	Nature of data
5.3(b), 5.3(c)	Processing environment
5.5	Measures to prevent re-identification of data holders/subjects
5.9	Third party recipients
6.2	Fees
8.2	Data format
8.2	Data size
9	Procedure to request reuse

7.1.2 Registers and records of altruistic and intermediation entities

DGA is also the first of its kind to regulate the activity of data intermediation service providers and altruistic organisations, with the requirements being outlined in Chapters III and IV. Regarding data intermediation, such entities *"aim[s] to establish commercial relationships for the purposes of data sharing between an undetermined number of data subjects and data holders on the one hand and data users on the other, through technical, legal or other means"* [2022g]. Moreover, as defined in Article 11 [2022g], entities wishing to provide data intermediation services must notify their competent national authority of their intentions (indicated by the (F) arrow in Figure 7.1). Subsequently, the national authority is mandated to publish and maintain an updated public register of intermediaries (represented by the (J) arrow) and oversee their activities ((I) arrow). Article 12 [2022g] includes a list of conditions for the provision of this service, e.g., providers should have tools to convert data into specific formats, use standards to promote interoperability across sectors, gather data

subjects' consent and data holders' permission terms, as well as update or withdraw these terms, maintain records of their activity, and appoint a legal representative if the data intermediation entity is not established in the EU (shown by the (G) arrow). Furthermore, the data intermediation service provider must maintain a log record of its activities (indicated by the (K) arrow). This log must include entity-related information available in the public register of intermediation providers, such as name, public website, legal status, ownership structure, subsidiaries, registration number, address, and details regarding the provided service type. A detailed list of the information used to notify authorities regarding the start of activity of a data intermediation service provider and the conditions to provide such service are available in Table 7.2.

For entities aiming to establish themselves as data altruism organisations, as outlined in DGA's Article 19 [2022g], the process entails submitting an application to their competent national authority (which may be the same authority regulating national data intermediation service providers for instance) to express their intentions (illustrated by the (M) arrow in Figure 7.1). Upon approval, the national authority is required to include information about the organisation in a public register of data altruism organisations (depicted by the (P) arrow). This register encompasses details such as the organisation's name, public website, legal status, form, registration number, contact information for the entity and its representative (if applicable), as well as information regarding the altruistic purposes underlying the organisation's activities. Furthermore, the organisation must publish and regularly update a uniform and structured record of its data altruism activities ((O) arrow). This record is submitted annually to the national authority for compliance verification (indicated by the (M) and (N) arrows). It must document the organisation's data altruism activities, including the nature and categories of data it handles. Additionally, it should contain logs of data users, their contact details, the processing date and duration, the altruistic purposes for data usage, fees paid by users or other sources of income, used technical processing means, and a summary of processing results. A detailed list of the information used to notify authorities regarding the start of activity of a data altruism organisation and the mandatory information items that need to be kept in the record of data altruism activity are available in Table 7.3.

7.1.3 Data altruism forms

Data altruism as a term was first introduced by the DGA. This term relates to the sharing of personal and non-personal data, based on data subjects' consent and data holders' permission, for the 'common good'. This entails purposes such as improving healthcare, combating climate change, or performing scientific research. Moreover, the proposed visions for a European Health Data Space [2022f] and the Data Act [2023] also emphasise the altruistic reuse of data, with the Health Data Spaces proposal, in particular, focusing on the challenges brought on by the access and sharing of electronic health data for scientific research and public interest purposes. In this context, each EU member state can establish its altruism policy and has to appoint a competent authority to oversee the activity of altruistic organisations (it can be the same as the one that supervises intermediation providers). As mentioned in the previously Section, said authority also has to keep up-to-date, public registers of altruistic organisations, and organisations themselves have to maintain records of their activities, in particular, to produce annual reports to share with the relevant competent authority. To facilitate this activity, a European data altruism consent form will be facilitated by the EC to *"allow the collection of consent or permission across Member States in*

Table 7.2: Information used to notify authorities regarding the start of activity of a data intermediation service provider and the conditions to provide such service, which need to be kept in a record of activities. Information items marked with (*) are kept in the public register of data intermediaries.

Article	**Information items**
11.3	Legal representative if not established in the EU (*)
11.6(a)	Name of the data intermediation services provider (*)
11.6(b)	Data intermediation services provider's legal status (*)
11.6(b)	Data intermediation services provider's form (*)
11.6(b)	Data intermediation services provider's ownership structure (*)
11.6(b)	Data intermediation services provider's relevant subsidiaries (*)
11.6(b)	Data intermediation services provider's registration number (*)
11.6(c)	Address of the data intermediation services provider (*)
11.6(c)	Address of the representative of the data intermediation services provider (*)
11.6.(d)	Public website (*)
11.6(e)	Data intermediation services provider's contact persons
11.6(e)	Data intermediation services provider's contact details
11.6(f)	Description of the data intermediation service (*)
11.6(f)	Type of data intermediation service (*)
11.6(g)	Estimated date for starting the activity (*)
11.6(g)	Date of the notification (*)
Article	**Conditions to provide intermediation services**
12(b)	Pricing
12(c)	Date and time of creation of data
12(c)	Geolocation data
12(c)	Duration of the activity
12(c)	Connections to other natural or legal persons established by the person who uses the data intermediation service
12(d)	Data format
12(d)	Convert the data into specific formats
12(e)	Tools to facilitate the exchange of data
12(f)	Terms of service
12(g,h,i,j,l)	Technical and organisational measures to protect, provision and ensure interoperability of data
12(k)	Notification of unauthorised transfer, access or use of the non-personal data
12(m)	Tools to facilitate exercising of data subjects' rights
12(n)	Third country jurisdiction
12(n)	Tools to obtain/withdraw consent/permission
12(o)	Log record of the data intermediation activity

Table 7.3: Information used to open activity as a data altruism organisation and respective information items that need to be kept in the record of said activity. Information items marked with (*) are kept in the public register of data altruism organisations.

Article	**Information items**
19.3	Legal representative if not established in the EU (*)
19.4(a)	Name of the entity (*)
19.4(b)	Entity's legal status (*)
19.4(b)	Entity's form (*)
19.4(b)	Entity's registration number (*)
19.4(c)	Statutes of the entity
19.4(d)	Entity's sources of income
19.4(e)	Address of the entity
19.4(e)	Address of the representative of the entity
19.4(f)	Public website (*)
19.4(g)	Entity's contact persons (*)
19.4(g)	Entity's contact details (*)
19.4(h)	Altruistic purposes (*)
19.4(i)	Nature of data
19.4(i)	Categories of personal data
Article	**Records of Altruism Activity**
20.1(a), 20.2(c)	Data users
20.1(a)	Data users' contact details
20.1(b)	Date of the processing of data
20.1(b)	Duration of the processing of data
20.1(c), 20.2(c)	Altruistic purpose for processing
20.1(d)	Fees paid by data users
20.2(a)	Information on its activities
20.2(b)	Description of the objectives of general interest
20.2(c)	Technical means used for processing
20.2(d)	Summary of the results of the data processing
20.2(e)	Sources of revenue

a uniform format", as described in Article 25 [2022g] and indicated by the (R) arrow in Figure 7.1. The development of this form involves consultation with GDPR's overseeing body, the EDPB, the forthcoming EDIB, and other relevant stakeholders. Once developed, it is to be adopted by data altruism organisations (depicted by the (S) arrow) to document both the consent provided by data subjects to share their personal data and the permissions granted by data holders to share their non-personal data. Furthermore, these forms are required to be maintained in both human and machine-readable formats.

7.2 Extending W3C vocabularies to cover DGA requirements

As an active contributor in the realm of data protection, the Semantic Web community possesses significant potential to aid with the compliance processes that such a legislation involve. This potential is based on the opportunities for interoperability that such a Web of Linked Data can support. In this context, Semantic Web technologies can be utilised:

- to model conditions for the reuse of public data;
- by data subjects, data holders, and data users to declare data access and usage policies in a machine-readable format; and
- by organisations and service providers to fulfil their legal obligations such as maintaining records of the processing activities.

Hence, vocabularies like the DPV, ODRL, and DCAT play a pivotal role in these procedures, serving as interoperable frameworks for the expression of legally-aligned documentation. DCAT and ODRL (including the work developed in OAC) can be used to publish records of activities as data catalogs and conditions for data access and usage as digital policies, respectively. Meanwhile, DPV provides a comprehensive, openly accessible set of taxonomies for articulating machine-readable metadata regarding the handling and usage of personal data. Using such solutions facilitates a transition from manual processes to automated ones by leveraging semantic technologies to uphold accuracy and scalability within the data-sharing ecosystems promoted by the DGA. In the forthcoming Sections, terms from the previously mentioned standards and specifications are employed to represent some of the information items delineated in Section 7.1. Additionally, the terms that cannot be expressed with existing vocabularies are provided in an open-source ad-hoc vocabulary – DGAterms. Examples to illustrate the practical applicability of the existing and developed semantic vocabularies are also supplied.

7.2.1 Policies for the reuse and sharing of public data

In Section 7.1.1, it is highlighted that public sector bodies must provide single information point providers with details on their data resources and conditions for their re-usage. This enables the providers to compile and maintain a searchable data asset list, facilitating data users' search and request for datasets for reuse. Table 7.4 presents terms from DPV, DCAT, OAC, and DCMI Metadata Terms that can be repurposed to model some of the concepts identified in Table 7.1.

Expanding upon the existing `dpv:LegalEntity` term, four new classes of entities were introduced in DGAterms, as subclasses, to represent data users, public sector bodies, compe-

Table 7.4: Information items that need to be represented to detail policies concerning the conditions of reuse of public sector bodies' datasets and respective reusable concepts from existing vocabularies.

Article	Information items	Terms from existing vocabularies
5.1	Public sector body information	`dpv:hasName`, `dpv:hasContact`
5.1	Competent body information	`dpv:hasName`, `dpv:hasContact`
5.2	Categories of data	`dpv:hasData`, `dpv:Data`
5.2	Purposes for usage and access	`dpv:hasPurpose`, `oac:Purpose`
5.3(a)	Nature of data	`dpv:hasData`, `dpv:AnonymisedData`, `dpv:PseudonymisedData`
5.3(b), 5.3(c)	Processing environment	`dpv:ProcessingContext`, `dpv:hasLocation` `dpv:WithinVirtualEnvironment`, `dpv:WithinPhysicalEnvironment`
5.5	Technical and operational measures to prevent re-identification of data holders/subjects	`dpv:Deidentification`
5.9	Third party recipients	`dpv:ThirdParty`
8.2	Data format	`dcat:mediaType`, `dcterms:format`
8.2	Data size	`dcterms:extent`

tent bodies, and single information point providers. These classes are named `DataUser`, `PublicSectorBody`, `DataReuseCompetentBody`, and `SingleInformationPointProvider`, respectively.

Additionally, subclasses of `SingleInformationPointProvider` were created to represent EU, national, regional, local, and sectorial-level single information point providers, as outlined in DGA's Article 8 [2022g]. Furthermore, to classify the nature of data kept by public sector bodies, as detailed in Article 3.1 [2022g], four new subclasses of textttdpv:Data – `ConfidentialData`, `CommerciallyConfidentialData`, `StatisticallyConfidentialData` and `IntellectualPropertyData` – were modelled in DGAterms as well. These subclasses represent data protected through `CommercialConfidentialityAgreements` or `StatisticalConfidentialityAgreements`, as well as data protected by intellectual property rights.

Beyond the aforementioned additions, the following concepts were included in the DGAterms vocabulary:

- `A5-9` for permissions to transfer, `A5-11` for model contractual clauses, and `A5-12` for adequacy decisions were added as subclasses of `dpv:DataTransferLegalBasis`.
- `DataReusePolicy` was introduced as a subclass of DPV's policy concept, and `DataTransferNotice` and `ThirdCountryDataRequestNotice` as subclasses of DPV's notice concepts. These represent the conditions for data reuse and the notices provided to data owners.

- `DataAssetList` and `DataReuseRequestProcedure` were modelled as subclasses of `dpv:OrganisationalMeasure` to represent the searchable asset list maintained by `SingleInformationPointProviders` and the procedure to request datasets, respectively.

To illustrate the application of both established and newly introduced terms, an instance of a `DataReusePolicy` concerning the dataset located at `http://example.com/dataset_001` is presented in Listing 7.1. This policy delineates the terms of usage for the dataset, specifying that it can be reused until the end of 2024, more specifically for the purpose of `ScientificResearch`. It is noteworthy that this policy, structured as an ODRL offer, outlines the conditions for utilising the dataset without granting any privileges to the data user. Single information point providers can utilise this policy to maintain an updated catalog of available assets along with their respective usage conditions. Furthermore, Listing 7.2 offers an example of a `DataAssetList` produced by a `SingleInformationPointProvider`, employing both pre-existing and newly devised terms. This list includes the aforementioned dataset, `http://example.com/dataset_001`, supplemented with additional metadata encompassing its data type, the governing policy (`http://example.com/offer_publicsectorbody`), data format and size, and any associated fees charged by the dataset publisher.

Listing 7.1 Data reuse policy, set by Public Sector Body X, that allows the reuse of a dataset until the end of 2024 for scientific research.

```
ex:offer_publicsectorbody a odrl:Offer, dgaterms:DataReusePolicy ;
    odrl:uid ex:offer_publicsectorbody ;
    odrl:profile oac: ;
    odrl:permission [
        odrl:target ex:dataset_001 ;
        odrl:action dgaterms:Reuse ;
        odrl:assigner ex:publicsectorbodyX ;
        odrl:constraint [
            odrl:and [
                odrl:leftOperand odrl:dateTime ;
                odrl:operator odrl:lteq ;
                odrl:rightOperand "2024-12-31"^^xsd:date ], [
                odrl:leftOperand oac:Purpose ;
                odrl:operator odrl:isA ;
                odrl:rightOperand dgaterms:ScientificResearch ] ] ] .
ex:publicsectorbodyX a dgaterms:PublicSectorBody ;
    dpv:hasName "Public Sector Body X" ;
    dpv:hasContact "mailto:publicsectorbodyX@email.com" ;
    dgaterms:hasCompetentBody [
        a dgaterms:DataReuseCompetentBody ;
        dpv:hasName "Competent Body X" ;
        dpv:hasContact "mailto:competentbodyX@email.com" ] .
```

7.2.2 Querying DGA-mandated public registers

As outlined in Section 7.1.2, data intermediation service providers and data altruism organisations are required to provide detailed information about the purpose of their operations to a public

Listing 7.2 Data asset list maintained by the Single Information Point Provider A.

```
ex:SIPPA_assets a dgaterms:DataAssetList, dcat:Catalog ;
    dcterms:description "Asset list maintained by SIPPA" ;
    dcterms:created "2023-12-04"^^xsd:date ;
    dcterms:publisher ex:SIPPA ;
    dcat:dataset ex:dataset_001 .
ex:SIPPA a dgaterms:SingleInformationPointProvider .
ex:dataset_001 a dcat:Dataset ;
    dcterms:publisher ex:publicsectorbodyX ;
    dpv:hasData dgaterms:StatisticallyConfidentialData ;
    dcterms:description "Dataset with statistically confidential data" ;
    dcterms:created "2023-12-04"^^xsd:date ;
    odrl:hasPolicy ex:policy_001 ;
    dgaterms:hasFee "0€"^^xsd:string ;
    dcat:mediaType <https://iana.org/assignments/media-types/text/csv> ;
    dcterms:extent "5.6MB"^^xsd:string .
```

register dedicated to such entities. This facilitates the creation of a database comprising information about these entities, which can be accessed by data users, holders, or subjects for retrieving or publishing data, e.g., for altruistic endeavors.

From the information items that need to be represented in a public register of data intermediation service providers, as specified in Table 7.2, 7 out of 14 can be defined using DPV, DCAT, and DCMI Metadata Terms declared in Table 7.5. In addition, from the 15 identified items that are necessary for a data intermediary to provide such service, 7 can also be described with DPV, DCAT, and DCMI Metadata Terms declared in Table 7.5.

As for the data altruism organisations, from the information items that need to be represented in a public register of such organisations, as specified in Table 7.3, 5 out of 9 can be defined using the DPV and DCAT terms declared in Table 7.6. In addition, from the 11 identified items which are necessary for a data altruism organisation to keep records of its activity, 5 can also be described with the DPV and DCMI Metadata Terms declared in Table 7.6.

In addition to these terms that can be reused from existing standards and specifications, DGAterms models the `DataIntermediationServiceProvider` concept to specify a data intermediation service provider (as a subclass of `dpv:LegalEntity`), and three new classes of entities to represent distinct types of intermediaries, i.e., `DataCooperativeServiceProvider`, `DataIntermediationServiceProviderForDataHolder`, and `DataIntermediationServiceProviderForDataSubject`.

The `DataAltruismOrganisation` concept is also introduced in DGAterms as a subclass of `dpv:NonProfitOrganisation`. Details concerning the nature of the entity, as outlined in Article 11.6(b) [2022g] regarding its legal status, form, ownership structure, subsidiary relationships, and registration number, fall beyond the scope of DGAterms as they pertain to organisational-specific items. However, as future work, upper ontologies like GIST [2023] or Schema.org [Guha et al., 2015] can be examined and potentially extended to encompass these concepts if deemed necessary.

Table 7.5: Information items that need to be kept to register the activity of data intermediation service providers and respective terms from existing vocabularies that can be reused.

Article	Information items	Terms from existing vocabularies
11.3	Legal representative if not established in the EU	`dpv:hasRepresentative`
11.6(a)	Name of the data intermediation services provider	`dpv:hasName`
11.6(c)	Address of the data intermediation services provider	`dpv:hasAddress`
11.6(c)	Address of the representative of the data intermediation services provider	`dpv:hasAddress`
11.6.(d)	Public website	`dcat:landingPage`
11.6(e)	Data intermediation services provider's contact persons	`dcat:contactPoint`
11.6(e)	Data intermediation services provider's contact details	`dpv:hasContact`
11.6(f)	Description of the data intermediation service	`dcterms:description`
11.6(g)	Date of the notification	`dcterms:issued`
Article	**Conditions to provide intermediation services**	**Terms from existing vocabularies**
12(c)	Date and time of creation of data	`dcterms:created`
12(c)	Geolocation data	`dpv:hasLocation`
12(c)	Duration of the activity	`dpv:hasDuration`
12(d)	Data format	`dcat:mediaType`, `dcterms:format`
12(g,h,i,j,l)	Technical and organisational measures to protect, provision and ensure interoperability of data	`dpv:hasTechnicalOrganisationalMeasure`
12(k)	Notification of unauthorised transfer, access or use of the non-personal data	`dpv:IncidentReportingCommunication`
12(m)	Tools to facilitate exercising of data subjects' rights	`dpv:isExercisedAt`

Table 7.6: Information items that need to be kept to record the activity of data altruism organisations and respective terms from existing vocabularies that can be reused.

Article	Information items	Terms from existing vocabularies
19.3	Legal representative if not established in the EU	`dpv:hasRepresentative`
19.4(a)	Name of the entity	`dpv:hasName`
19.4(e)	Address of the entity	`dpv:hasAddress`
19.4(e)	Address of the representative of the entity	`dpv:hasAddress`
19.4(f)	Public website	`dcat:landingPage`
19.4(g)	Entity's contact persons	`dcat:contactPoint`
19.4(g)	Entity's contact details	`dpv:hasContact`
19.4(i)	Nature of data	`dpv:hasData`
19.4(i)	Categories of personal data	`dpv:hasPersonalData`, pd extension
Article	**Records of altruism activity**	**Terms from existing vocabularies**
20.1(a)	Data users' contact details	`dpv:hasContact`
20.1(b)	Duration of the processing of data	`dpv:hasDuration`
20.2(b)	Description of the objectives of general interest	`dcterms:description`
20.2(c)	Technical means used for processing	`dpv:hasTechnicalMeasure`, `dpv:PrivacyPreservingProtocol`
20.2(d)	Summary of the results of the data processing	`dpv:hasOutcome`

Moreover, a `PublicRegister` class was also modelled, as a new subclass of DPV's organisational measures, and its respective subclasses `PublicRegisterOfDataIntermediationServiceProviders` and `PublicRegisterOfDataAltruismOrganisations` to represent public registers of data intermediaries and altruistic organisations, respectively.

An illustration of a register of data intermediation service providers, incorporating both existing and newly introduced terms, is available in Listing 7.3. The register `ex:publicregistry_DI_PT` presents a comprehensive list of intermediaries operating in Portugal. In addition to storing metadata concerning the national authority `ex:nationalauthority_PT` and creation dates, the register already includes a registered company, `ex:DISP_Y`, which operates as a `DataCooperative`. DPV's `hasName`, `hasContact`, and `hasAddress`, along with DCAT's `landingPage`, are used to detail the providers offering data intermediation services. On the other hand, DCMI's `description`, `created`, and `publisher` predicates are essential in describing metadata pertaining to the register itself.

Storing the public registers in RDF format, utilising existing and crafted semantic vocabularies such as DGAterms, facilitates seamless querying of these structures, using a language like SPARQL, for automated retrieval of information concerning data intermediation service providers (or of data altruism organisations). An example query targeting data cooperatives is presented in Listing 7.4,

Listing 7.3 Example of a public register of data intermediation service providers.

```
ex:publicregistry_DI_PT a
 ↪  dgaterms:RegisterOfDataIntermediationServiceProviders ;
    dcterms:description "Public register of intermediaries in PT" ;
    dcterms:created "2023-12-15"^^xsd:date ;
    dcterms:modified "2023-12-23"^^xsd:date ;
    dcterms:publisher ex:nationalauthority_PT ;
    dgaterms:hasDataIntermediationServiceProvider ex:DISP_Y .
ex:nationalauthority_PT a dgaterms:DataIntermediationAuthority ;
    dpv:hasName "Data Intermediation Authority of Portugal" ;
    dpv:hasContact "mailto:nationalauthority_PT@email.com" ;
    dpv:hasJurisdiction "PT" .
ex:DISP_Y a dgaterms:DataCooperative ;
    dpv:hasName "Data Cooperative Y" ;
    dpv:hasAddress "Lisboa, Portugal" ;
    dcterms:description "Provider of anonymised geolocation data" ;
    dcat:landingPage <http://cooperativeA.com/> ;
    dcterms:date "2023-12-23"^^xsd:date .
```

which returns the providers offering such services, along with their names and public websites.

Listing 7.4 SPARQL query to retrieve data cooperatives.

```
SELECT DISTINCT ?Provider ?Name ?Web WHERE {
    ?Provider a dgaterms:DataCooperative .
    ?Provider dpv:hasName ?Name .
    ?Provider dcat:landingPage ?Web . }
```

Furthermore, beyond public registers, as mentioned in Section 7.1.2, intermediaries and altruism organisations must keep records of their respective activities for the competent authorities to verify their compliance with DGA-mandated requirements for providing such services. In Listing 7.5, a representation of a record of data altruism activity is presented, using the modelled `RecordOfDataIAltruismActivity`. These activity logs should be linked with the entities utilising the data and can be documented utilising DPV's `hasPersonalDataHandling`. This property can be used to supply details regarding the data processing, encompassing aspects such as duration, purpose, and categories of (personal) data involved.

7.2.3 Uniform data altruism forms

As demonstrated by the examples presented in the preceding Sections, the identified vocabularies, along with the one developed in this Chapter, are suitable for automating the generation of compliance documentation for single information point providers, data altruism organisations, and data intermediation service providers alike. Following these steps, the representation of consent forms for data subjects and permission forms for data holders should also be possible to automate with such technologies. Through their (re)use, the DGA-mandated European data altruism forms can be ensured to be interoperable across the EU. An instance of a consent form, utilising the DGAterms-modelled `EuropeanDataAltruismConsentForm`, completed by

Listing 7.5 Example of a record of data altruism activity logs.

```
ex:altruism_logs a dgaterms:RecordOfDataIAltruismActivity ;
    dcterms:description "Logs of Data Altruism Organisation A" ;
    dcterms:created "2023-11-04"^^xsd:date ;
    dcterms:modified "2023-11-13"^^xsd:date ;
    dcterms:publisher ex:altruism_A ;
    dcat:record ex:log_001 .
ex:altruism_A a dgaterms:DataAltruismOrganisation ;
    dpv:hasName "Data Altruism Organisation A" ;
    dpv:hasAddress "Lisboa, Portugal" ;
    dcat:landingPage <http://example.com/altruism_A> .
ex:log_001 a dcat:CatalogRecord ;
    dcterms:created "2023-11-13"^^xsd:date ;
    dgaterms:hasDataUser ex:userZ ;
    dgaterms:hasFee "1000€"^^xsd:string ;
    dpv:hasPersonalDataHandling [
        dcterms:description "Download and reuse anonymised health records
        ↪ to improve healthcare" ;
        dpv:hasProcessing dgaterms:Download,dgaterms:Reuse ;
        dpv:hasDuration 6226453 ;
        dpv:hasPurpose dgaterms:DataAltruism,dgaterms:ImproveHealthcare ;
        dpv:hasPersonalData pd:HealthRecord ;
        dpv:hasTechnicalMeasure dpv:Anonymisation ] .
ex:userZ a dgaterms:DataUser ;
    dpv:hasName "Data User Z" ;
    dpv:hasContact "mailto:user_z@email.com" .
```

a data subject, is detailed in Listing 7.6. In this example, an altruistic purpose for processing is used. To this end, a taxonomy of `DataAltruism` concepts, modelled from DGA's Article 2.16 [2022g], are presented in DGAterms, with seven new purposes applicable to a data altruism context: `ImproveHealthcare`, `CombatClimateChange`, `ImproveTransportMobility` (utilised in Listing 7.6), `ProvideOfficialStatistics`, `ImprovePublicServices`, `ScientificResearch`, and `PublicPolicyMaking`. Furthermore, additional purposes referenced throughout the DGA are also incorporated into the vocabulary developed in this Chapter.

Similarly, Listing 7.7 presents an example of a permission form issued by a data holder.

7.3 DGAterms development and evaluation

The motivation and identified requirements for the development of the DGAterms, which were outlines in the previous Sections, are consolidated in the vocabularies's ORSD available in Table 7.7.

The methodology followed to produce the vocabulary is described in Section 3.7.2. The vocabulary human-readable documentation and machine-readable file are available at https://w3id.org/dgaterms using content negotiation. The HTML documentation includes a description of the terms defined in DGAterms, which was conducted and validated with domain experts, diagrams with graphical representations of the several taxonomies included in the vocabulary,

Table 7.7: Ontology Requirement Specification Document of DGAterms.

DGAterms – Vocabulary to describe information flows in the Data Governance Act
1. Purpose
The purpose of DGAterms is to provide concepts to describe conditions for the reuse of data held by public sector bodies, to publish and maintain public registers of entities and records of their activities, and to record data altruism consent and permission forms.
2. Scope
The scope of this vocabulary is limited to the flows of information depicted in DGA's Chapters II, III, and IV, and respective records of information that must be maintained by DGA entities in order to fulfil their obligations. DGAterms also promotes the usage of ODRL, OAC, DPV, DCAT, and DCMI Metadata Terms to model policies for the reuse of personal and non-personal data, to keep catalogs of entities, and to represent consent and permission terms.
3. Implementation Language
RDF, RDFS
4. Intended End-Users
Providers of single information points, data intermediation and altruism services, and maintainers of public registers of entities.
5. Intended Uses
Use 1. Describing entities, and processes involved in the sharing of data compliant with the DGA. Use 2. Expressing information regarding the representation of data-sharing policies and consent terms. Use 3. Generating records of altruistic and data intermediary activities that can be audited by national and European authorities.
6. Ontology Requirements
a. Non-Functional Requirements
NFR 1. The ontology is published online with HTML documentation, following W3C's specification format.
b. Functional Requirements: Groups of Competency Questions
CQD1. Which legal basis can be used for the processing of personal and non-personal data regulated by the DGA? CQD2. Which processing operations can be performed on the data? CQD3. Who are the entities involved in DGA information flows? CQD4. Which purposes are considered altruistic under the DGA? CQD5. Which technical and organisational measures can be used to protect, provide, and ensure interoperability of data? CQD6. Which data types can be shared by public sector bodies? CQD7. What kinds of notices should be provided by DGA-related entities? CQD8. Which public registers should be maintained by authorities? CQD9. What rights are provided to data subjects and data holders in the context of DGA?

Listing 7.6 Data altruism consent form, where the data subject consents to the usage of their location data for improving mobility.

```
ex:consentForm_001 a dgaterms:EuropeanDataAltruismConsentForm ;
    dpv:hasIdentifier <http://example.com/consentForm_001> ;
    dpv:hasDataSubject ex:Anne ;
    dpv:isIndicatedBy ex:Anne ;
    dpv:isIndicatedAtTime "2022-12-14"^^xsd:date ;
    dpv:hasPersonalDataHandling [
        dpv:hasPurpose dgaterms:DataAltruism,
         ↪ dgaterms:ImproveTransportMobility ;
        dpv:hasLegalBasis eu-gdpr:A6-1-a ;
        dpv:hasPersonalData pd:Location ;
        dpv:hasProcessing dpv:Use, dpv:Store ;
        dpv:hasDataController [
            a dpv:DataController, dgaterms:DataAltruismOrganisation ;
            dpv:hasName "Company A" ] ] .
```

Listing 7.7 Permission for data altruism where data holder A allows the usage of their anonymised data for the purpose of providing official statistics.

```
ex:permissionForm_001 a dpv:Permission ;
    dpv:hasIdentifier <http://example.com/permissionForm_001> ;
    dgaterms:hasDataHolder ex:dataHolderA ;
    dpv:isIndicatedBy ex:dataHolderA ;
    dpv:isIndicatedAtTime "2022-12-15"^^xsd:date ;
    dpv:hasPersonalDataHandling [
        dpv:hasPurpose dgaterms:DataAltruism,
         ↪ dgaterms:ProvideOfficialStatistics ;
        dpv:hasLegalBasis dgaterms:A2-6 ;
        dpv:hasData dpv:AnonymisedData ;
        dpv:hasProcessing dpv:Use, dpv:Store ;
        dpv:hasDataController [
            a dpv:DataController, dgaterms:DataAltruismOrganisation ;
            dpv:hasName "Company A" ] ] .
```

RDF examples of policies, registers of entities, and records of activities that use DGAterms terms. The vocabulary documentation also includes metadata, such as the identity of the creators and publishers of the vocabulary, the dates of creation and last modification, or the version number.

The source code is hosted at https://w3id.org/dgaterms/repo, under the CC-BY-4.0 license. The repository can also be used by DGAterms implementers to suggest new inclusions to the vocabulary and to report bugs through GitHub Issues.

In terms of quality evaluation, the OOPS! tool was used to detect common errors in ontology development, such as missing domain or range properties or missing human-readable annotations. No critical nor important issues were detected through this evaluation. Moreover, FOOPS! was used to evaluate the alignment of the developed vocabularies with the FAIR principles. The following results were obtained:

- Findable – 8/9

- Accessible – 2/3
- Interoperable – 2/3
- Reusable – 8.83/9
- FOOPS! overall score – 91%

These outcomes are aligned with the scores obtained for OAC, PLASMA, and DUODRL. Furthermore, DGAterms obtained a good score in all FAIR aspects. In terms of improvements, DGAterms can be submitted to LOV to be recorded in a public registry of ontologies. This will improve both the findability and the accessibility of the vocabulary.

7.4 SoDA – Solid for Data Altruism

This Section features an architecture designed to enable data altruism as a service, utilising the Solid protocol and ODRL policies to facilitate the sharing of personal data for altruistic purposes in a manner that respects data protection principles. The policies are articulated using OAC and the DGAterms concepts related to data altruism. Furthermore, the Solid Data Altruism application, SoDA, is introduced, which allows (a) individuals to create policies for sharing their personal data for altruistic purposes, (b) data users to request access to datasets for altruistic purposes, and (c) data altruism organisations to manage metadata concerning available datasets.

7.4.1 Solid architecture for data altruism

The diagram presented in Figure 7.2 provides a broad summary of an architecture designed to implement data altruism as a service through the usage of the Solid protocol, with the central component being a Solid for Data Altruism application, known as SoDA. The objective of this architecture is to start a proof of concept decentralised ecosystem for data altruism, emphasising the capability of data subjects to share personal data and data users to discover available datasets suitable for altruistic purposes, in line with data protection principles in the EU as the information disclosed about each dataset is limited to its data type and the intended purpose for its utilisation.

As previously described, within a Solid-based architecture, users are recognised through a WebID and utilise Solid Pods to either store data or request access to stored data, adhering to the specifications outlined in the Solid protocol. When personal data resides within Pods, GDPR and DGA requirements come into effect, with individuals who store their personal data in Pods being categorised as 'data subjects'. Additionally, both data subjects and data users administer data access via Solid applications. In this setting, the SoDA application is introduced to:

(a) empower data subjects to create policies for sharing their personal data with altruistic intent;

(b) enable users to seek access to datasets based on their data type and intended purpose for usage; and

(c) facilitate organisations in offering data altruism services by maintaining metadata about accessible datasets within their own Solid Pod, without the necessity of storing the data themselves, aligning with Solid's decentralised principles.

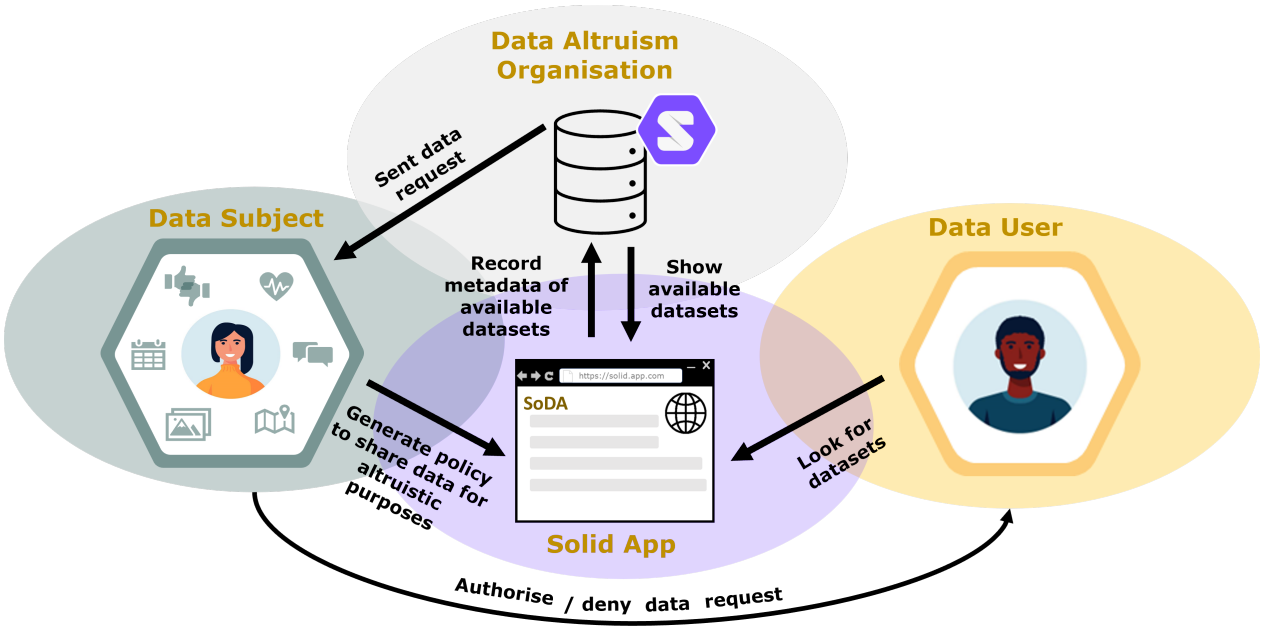


Figure 7.2: High-level diagram specifying an architecture to implement data altruism as a service using Solid and SoDA, a Solid application to edit policies/search for data for altruistic purposes, adapted from Esteves [2023].

With SoDA, individuals can craft data access policies governing their personal data, which are stored in their Solid Pod and can be shared with a data altruism organisation, which exclusively records metadata about the dataset and access conditions. These records are leveraged to present available datasets to data users, safeguarding the privacy of data subjects by revealing solely the data type and permissible purpose of use, while concealing their identity. Should users identify datasets of interest, the data altruism organisation serves as an intermediary, forwarding data requests to the data subjects for them to authorise or deny access to the requested data. Policies are modelled using OAC – and by consequence, ODRL and DPV –, and DGAterms, and the catalogues of datasets kept by the altruistic company are based on DCAT. Detailed instructions on how to install, launch, and use SoDA are available on the source code repository[1]. Figures 7.3 and 7.4 present a screenshot of SoDA's (i) policy editor UI and (ii) dataset request UI, respectively.

7.4.2 SoDA coverage, maintenance, and future work

SoDA is published and archived according to the methodology described in Section 3.7.3. Furthermore, SoDA's source code is hosted at https://w3id.org/people/besteves/soda/repo, under the CC-BY-4.0 license. Further information on this proof of concept application can be found at https://w3id.org/people/besteves/demo/iswc23, including a demonstration of the features of the app. The repository can also be used by SoDA users to suggest new features to be added to the app and to report bugs through GitHub Issues. Currently, SoDA's app coverage encloses terms from OAC/DPV taxonomies of purposes and personal data categories, focusing on the altruistic purposes described in the DGA. As future work, SoDA can

[1] https://w3id.org/people/besteves/soda/repo

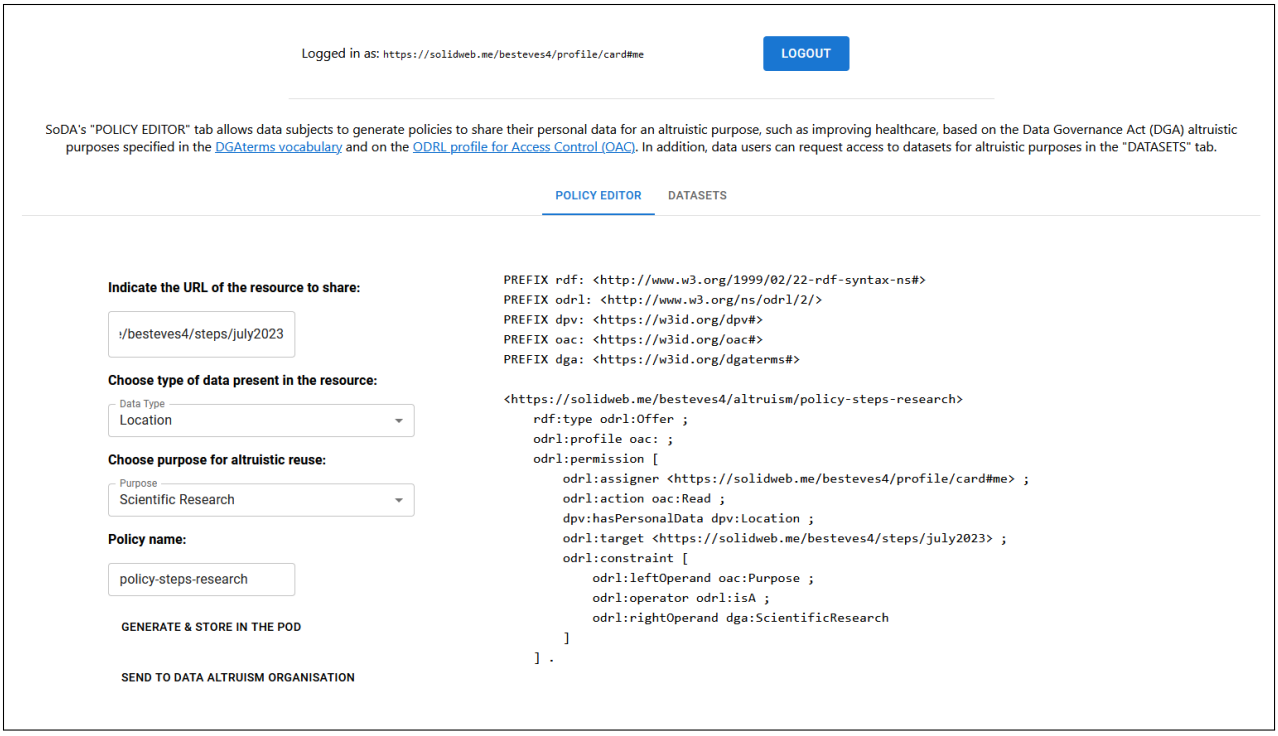


Figure 7.3: Screenshot of SoDA policy editor UI, which generates OAC and DGAterms-based policies.

be extended to include all terms present in the previously mentioned DPV's data taxonomies, as well as to cover all constraints defined in the OAC profile, e.g., restrict legal bases, recipients or specify the technical and organisational measures used by data controllers to ensure the secure processing of personal data. Additionally, user studies should be performed to assess the design choices included in the policy editor and dataset request UI, as well as to understand what type of additional controls people want to have on top of what is legally mandated, in particular, related to sharing data for altruistic purposes, e.g., temporal constraints or duties for the data user to fulfil after accessing the data.

7.5 Lessons learned for the (Personal) Data Spaces future

While the European data strategy is robust, it presents numerous interoperability challenges that must be overcome to establish shared data spaces across individuals, businesses, and governments. Consequently, the work developed in this Chapter, focusing on analysing the requirements of the DGA and developing a unified semantic model for documenting the activities of public sector bodies, intermediaries, and altruistic organisations, represent an initial stride towards addressing these interoperability hurdles. In this context, semantic technologies offer promising applications in operationalising compliance with the DGA. Among them, the following advantages can be mentioned:

- **Enhanced Interoperability** – Semantic technologies enable better integration and interoperability of data coming from distinct sources and systems, facilitating the work of data altruism organisations and intermediation providers in consolidating datasets coming from

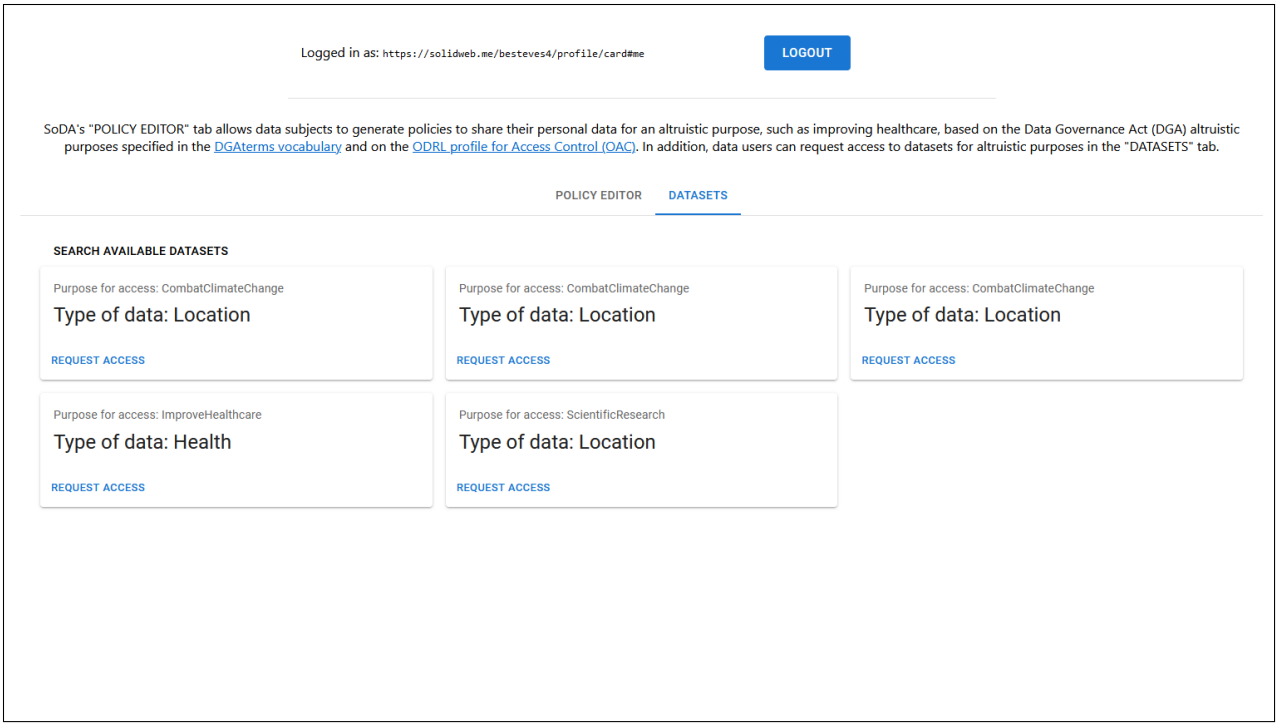


Figure 7.4: Screenshot of SoDA dataset request UI, which allows data users to request access to a dataset for a specific purpose.

different data subjects and other data holders.

- **Improved Knowledge Management** – By structuring, organising and publishing data with semantic standards, e.g., DCAT for cataloguing datasets, data discovery and analysis can be performed more efficiently.
- **Enhanced Decision Support** – Semantic technologies enable the development of Web agents with sophisticated decision support systems that can provide actionable insights to data subjects and data holders, aiding them in making informed decisions when it comes to the use of their data.
- **Improved Legal Support** – Using a common semantic model to tackle legal requirements from distinct data-related regulations, e.g., DPV, aids businesses to have a shared understanding of regulatory provisions and to comply with their legal duties related to the processing of personal and non-personal data.

While these advantages are promising, it should be acknowledged that there are challenges that need to be addressed to support the sustainable development of data altruism and data intermediation services towards having common European data spaces:

- Most constraints specified in the policies cannot be automatically enforced, and the declarative nature of the policies may inadvertently result in data misuse.
- If an agreement is not reached in terms of which semantic models need to be used, interoperability will be difficult to achieve among data subjects, holders, users and even public

authorities.

As future contributions, it is imperative to explore the potential of the Data Act and the European Health Data Space proposals to enhance the outreach of DPV and achieve the envisioned interoperability to have common European data spaces. Moreover, to complement the described system, future efforts should include: (i) implementing SHACL shapes to validate data reuse and data altruism policies, (ii) conducting usability tests to evaluate the design choices made in SoDA, including scalability testing, which may involve utilising data aggregators to manage organisations seeking simultaneous access to numerous datasets, (iii) enhancing/automating the process of authorising/denying data requests through technologies such as RDF surfaces [Hochstenbach et al., 2023] to perform reasoning tasks over data policies, and (iv) facilitating the creation of immutable agreements, e.g., by integrating Verifiable Credentials into the Solid ecosystem [Braun and Käfer, 2022a] to digitally sign data usage conditions, which can be utilised by authorities in case of misuse by data users.

Part IV

CONCLUSIONS

Chapter 8

Conclusions

In a world where AI-based technologies are taking over and distrust in data-consuming services is at its highest point, BigTech companies prefer to deal with the consequences of their unlawful practices than provide the necessary tools to data subjects to make the right decisions over the processing of their personal data. Endowed with enhanced interoperability and transparency features, the decentralised Semantic Web aims to aid data subjects in taking control of the publication and movement of their personal data. As such, legally-aligned vocabularies and services were produced to support policy-based access to data in decentralised settings, providing accountability and enhanced transparency to people looking at regaining trust in Web services. To this end, Section 8.1 concludes the Thesis with a discussion on the extent to which the research objectives have been fulfilled through the contributions described in Chapters 4 to 7. Section 8.2 provides lines of future work arising from the research presented within this Thesis. Finally, Section 8.3 discusses the scientific, technological, and societal impact of the Thesis' contributions.

8.1 Fulfilment of research objectives

In Section 3.5, the main research question driving the development of this Thesis is presented as *"Are Semantic Web vocabularies and decentralised technologies able to support the exercising of data subject rights and determine the access conditions to personal data?"*. As such, the following objective, *"Research methodologies and design vocabularies and services to aid EU data subjects in taking control of the movement of their personal data."*, was identified to guide the Thesis development towards answering such question. This objective was further divided into three sub-objectives and the extent of their fulfilment, based on the work presented in the previous Chapters and consolidated in the contributions outlined in Section 3.6, is discussed in this Section.

O1 – Assist entities in the expression of data protection-related information The work developed in Chapter 4, in particular in Sections 4.2 and 4.3, and Chapters 5 and 7 targeted the fulfilment of this objective, with OAC and PLASMA targeting the expression of information legally aligned with the GDPR and DGAterms with the recently applicable DGA. Moreover, the developed Solid-based UIs for policy generation, SOPE (Section 6.2.1) and SoDA (Section 7.4), can assist users with expressing their policies using said vocabularies.

O2 – Use machine-readable policies for accessing decentralised personal data The architecture and algorithms described in Chapter 6 targeted the achievement of this objective and, in particular, the proof of concept described in Section 6.3, including the work on DUODRL, showcases its fulfilment for a health data sharing scenario.

O3 – Aid the exercising of GDPR's data subject rights The work developed in Section 4.4, on having machine-readable information related to the exercising of data subject rights with DPV, and in Section 6.2.4, on having a service to assist in the exercising and recording of such rights exercise activities, aimed to fulfil this goal. Moreover, the work on OAC also enables the data subjects' right to be informed, as described in GDPR's Articles 13 and 14.

These results confirm the hypotheses drafted in Section 3.2 as Semantic Web technologies can be used to successfully express data protection-related information, including the definition of data subject's privacy preferences as access control policies related to their personal data. This is supported by the evaluation performed in Section 4.5 and the legal analysis described in Chapter 5. Furthermore, said technologies can be used to increase the transparency and accountability of decentralised data environments, in particular when it comes to the involved entities and infrastructure, as currently systems such as Solid do not keep provenance metadata regarding the providers of storage, applications or other services, neither do they keep logs of the activities of the involved actors or processes, e.g., no records are kept of users updating existing resources or of changes to identity provider of the data subject. The developed vocabularies allow the expression of such information, which can be used by data subjects to inspect the usage of their data and external auditors to validate if personal data handling practices are done according to the law. Moreover, the improved access control mechanism, based on the proposed architecture and policy matching algorithm, provides alignment with the transparency information requirements outlined in the GDPR, extending Solid's read-write access control system with legally-aligned policies that allow the expression of purposes for accessing data, as well as legal grounds, processing operations and particular data types. By confirming the outlined hypotheses, the work in this Thesis leaves data subjects one step closer to having control over the publication and movement of their personal data, and in return data controllers with a variety of tools to express information related to their GDPR duties.

8.2 Future work

With the European Commission expanding the scope of its legislative initiatives from personal data to non-personal data and to the under-development common European data spaces, further opportunities for future work based on the contributions proposed in this Thesis can be envisioned.

Beyond consent The contributions proposed in this Thesis focus on the usage of consent as the legal basis for the processing of personal data. This places a lot of responsibilities on the data subject, which would need to individually approve every request for data. To avoid consent fatigue, other GDPR legal bases, described in Article 6 [2016b], should also be used by data controllers for the lawful processing of personal data, in particular the usage of contracts or legitimate interests. By using such legal grounds automation would be possible, lessening the burden on data subjects.

As such, the proposed work on policies can be used by software agents to automatically provide access to data under a legal basis which is not consent.

Delegation and data reuse The central theme of this Thesis revolves around empowering data subjects to exert control over the fate of their data. However, there are situations where data subjects might want to rely on other people or organisations to make the decisions for them. A first step was taken by the work on DGAterms, which can be used by data subjects to specify under which conditions their data is available, and in turn intermediation service providers and altruistic organisations make it available to data users, following these conditions. This work should be expanded to cover a wider set of use cases, e.g., allow data subjects to delegate the decision of what happens with their health data to their doctor. Moreover, the EHDS proposal [2022f] expands on DGA's altruistic intentions by providing an extensive set of altruistic purposes for the secondary use of data for improved healthcare services or innovative research on rare diseases.

Usage control and data spaces While this Thesis focused on building solutions to aid data subjects and data controllers in dealing with access to data, the envisioned European data spaces are putting the focus on usage control solutions. As proven by the work on DUODRL, the proposed vocabularies can be easily extended to include usage constraints, e.g., publishing the results of the research, and the proposed architecture should be developed to include a usage control enforcement component. Moreover, additional research should be performed on how to achieve GDPR compliance in data space environments where the 'accountability' and 'data protection by design' principles will be key to ensure a proper treatment of personal data. The Agencia Española de Protección de Datos [2023] provides a first approach to this, by suggesting the usage of *"various European standards, specific standards and vocabulary in the field of Data Spaces"*.

Web agents Building on the previous three points, decentralised data environments such as Solid can rely on Web agents to assist data subjects, data controllers, and newly-introduced DGA entities to exercise their rights or fulfil their duties in an automated manner. In this context, agents can be useful in making decisions for the data subjects, according to their preferences, help data controllers to compile the necessary compliance documentation, or aid altruistic organisations in their data meddling functions.

Improved user interfaces This Thesis showcases three proof of concept user interfaces for data subjects to edit their privacy preferences and exercise their right of access. These interfaces should be further improved to cover a wider range of policies, as well as to enclose tools to assist data subjects in the exercise of all of their GDPR data subjects and further rights from other European and non-European laws. Moreover, privacy dashboards for users to manage their data and understand how it is being used and by whom should also be developed.

Contextualised and verifiable data As described by Verborgh [2023], *"Data without context is meaningless; data without trust is useless."* When data subjects give access to their data, they want to know that it is going to be used according to their preferences and not for purposes that they do not agreed with. On the other hand, data controllers and data users want to know that they are receiving complete and correct data, while supervisory authorities need to have access

to contextualised data access and usage metadata to verify that it is not being misused. As such, to have trustful and responsible data flows in decentralised data environments, data should be accessed and shared with accompanying access and usage policies as well as contextual metadata and digital signatures.

Interaction of data protection and AI laws Beyond GDPR and the DGA, the European strategy for data also introduced the Digital Services Act (DSA) [2022c], the Digital Markets Act (DMA) [2022b], and the Data Act [2023] legislative initiatives. Moreover, the EC also launched the first-ever legal framework on AI, the AI Act [2021e]. Additionally, the rest of the world is following the European approach, with new data and AI-related legal frameworks being launched outside the EU. To deal with the new requirements brought on by these new laws, further vocabularies need to be developed and integrated into DPV's existing framework, which currently covers jurisdiction-agnostic as well as GDPR and DGA-specific terms. Additionally, the proposed ODRL agreement algorithm can be used to check whether AI model's deployers follow the intended purpose of their developers as mandated by the AI Act. Furthermore, a study of how these laws are related and in which data processing scenarios they apply still needs to be performed to be incorporated into the developed decentralised systems.

8.3 Impact

The outcomes of this Thesis highlight the advantages and possibilities of a joint technological and legal approach to personal data management. While past approaches have nearly exclusively focused on the technological prevention of legally undesirable behaviours, the scientific, technological, and societal solutions derived from this Thesis' work enable a much wider perspective on the problem space. The contributions of the Thesis are carefully supported and validated by legal research and also challenge this research domain, e.g., by highlighting the shortcomings of using consent as a legal ground.

As such, this work has the potential for a strong, interdisciplinary, scientific, technological, and societal impact, which can serve as the foundation of important techo-legal research to come. This statement is supported by the several published RDF vocabularies derived from this Thesis' contributions, which are finding adoption with industry practitioners. The active participation and contribution to W3C specification processes also further highlights the aforementioned impact in standardisation activities.

Bibliography

Constitución Española, 1978.

Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data. *Official Journal of the European Union L 281*, 1995.

Charter of Fundamental Rights of the European Union. *Official Journal of the European Communities*, 2000.

Directive 2002/58/EC of the European Parliament and of the Council of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector (Directive on privacy and electronic communications). *Official Journal of the European Union L 201*, pages 37–47, 2002.

Legea 506/2004 Privind Prelucrarea Datelor cu Caracter Personal si Protectia Vietii Private in Sectorul Comunicatiilor Electronice, 2004.

Break down these walls. *The Economist*, 2008a. ISSN 0013-0613. URL https://www.economist.com/leaders/2008/03/19/break-down-these-walls.

Real Decreto 1720/2007, de 21 de diciembre, por el que se aprueba el Reglamento de desarrollo de la Ley Orgánica 15/1999, de 13 de diciembre, de protección de datos de carácter personal. *Boletín Oficial del Estado 17*, 2008b.

Directive 2009/136/EC of the European Parliament and of the Council of 25 November 2009 amending Directive 2002/22/EC on universal service and users' rights relating to electronic communications networks and services, Directive 2002/58/EC concerning the processing of personal data and the protection of privacy in the electronic communications sector and Regulation (EC) No 2006/2004 on cooperation between national authorities responsible for the enforcement of consumer protection laws. *Official Journal of the European Union L 337*, pages 11–36, 2009.

Semantic Web Case Studies and Use Cases, 2012. URL https://www.w3.org/2001/sw/sweo/public/UseCases/.

Regulation (EU) No 910/2014 of the European Parliament and of the Council of 23 July 2014 on electronic identification and trust services for electronic transactions in the internal market and repealing Directive 1999/93/EC. *Official Journal of the European Union L 257*, pages 73—-114, 2014.

Directive (EU) 2016/680 of the European Parliament and of the Council of 27 April 2016 on

the protection of natural persons with regard to the processing of personal data by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and on the free movement of such data, and repealing Council Framework Decision 2008/977/JHA. *Official Journal of the European Union L 119*, 2016a.

Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). *Official Journal of the European Union L 119*, pages 1–88, 2016b.

Orange Romania SA v Autoritatea Naţională de Supraveghere a Prelucrării Datelor cu Caracter Personal (ANSPDCP), 2020.

Appendix B: Using the Data Privacy API, 2021a. URL https://bit.ly/3MPJFLn.

Proposal for a Regulation of the European Parliament and of the Council amending Regulation (EU) No 910/2014 as regards establishing a framework for a European Digital Identity. *COM/2021/281 final*, 2021b.

Janeiro Digital at Solid World: NHS Personal Health Stores with XFORM Health and Solid, 2021c. URL https://www.janeirodigital.com/blog/janeiro-digital-at-solid-world-nhs-personal-health-stores-with-xform-health-and-solid/.

PDS Interop, 2021d. URL https://pdsinterop.org/.

Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act). 2021e.

Commission staff working document on Common European Data Spaces, 2022a. URL https://ec.europa.eu/newsroom/dae/redirection/document/83562.

Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 September 2022 on contestable and fair markets in the digital sector and amending Directives (EU) 2019/1937 and (EU) 2020/1828 (Digital Markets Act). *Official Journal of the European Union L 265*, pages 1–66, 2022b.

Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act). *Official Journal of the European Union L 277*, pages 1–102, 2022c.

EDPB-EDPS Joint Opinion 03/2022 on the Proposal for a Regulation on the European Health Data Space, 2022d. URL https://edpb.europa.eu/system/files/2022-07/edpb_edps_jointopinion_202203_europeanhealthdataspace_en.pdf.

Implementing the OpenGDPR API, 2022e. URL https://support.appsflyer.com/hc/en-us/articles/11332840660625-OpenDSR-API.

Proposal for a Regulation of the European Parliament and of the Council on the European Health Data Space, 2022f.

Regulation (EU) 2022/868 of the European Parliament and of the Council of 30 May 2022 on European data governance and amending Regulation (EU) 2018/1724 (Data Governance Act). *Official Journal of the European Union L 152*, pages 1–44, 2022g.

Use the Microsoft Graph compliance and privacy APIs, 2022h. URL https://bit.ly/3MORtwZ.

Regulation (EU) 2023/2854 of the European Parliament and of the Council of 13 December 2023 on harmonised rules on fair access to and use of data and amending Regulation (EU) 2017/2394 and Directive (EU) 2020/1828 (Data Act). *Official Journal of the European Union L 2023/2854*, 2023.

gist ontology v12.0.0, 2023. URL https://w3id.org/semanticarts/ontology/gistCore.

Sushant Agarwal, Simon Steyskal, Franjo Antunovic, and Sabrina Kirrane. Legislative Compliance Assessment: Framework, Model and GDPR Instantiation. In Manel Medina, Andreas Mitrakas, Kai Rannenberg, Erich Schweighofer, and Nikolaos Tsouroulas, editors, *Privacy Technologies and Policy*, Lecture Notes in Computer Science, pages 131–149. Springer International Publishing, 2018. ISBN 978-3-030-02547-2. doi: 10.1007/978-3-030-02547-2_8.

Agencia Española de Protección de Datos. Approach to Data Spaces from GDPR Perspective, 2023. URL https://www.aepd.es/documento/approach-to-data-spaces-from-gdpr-perspective.pdf.

Rakesh Agrawal, Jerry Kiernan, Ramakrishnan Srikant, and Yirong Xu. XPref: a preference language for P3P. *Computer Networks*, 48(5):809–827, 2005-08. ISSN 13891286. doi: 10.1016/j.comnet.2005.01.004.

Ines Akaichi. Semantic Technology based Usage Control for Decentralized Systems, 2022. URL http://arxiv.org/abs/2206.04947.

Ines Akaichi, Giorgos Flouris, Irini Fundulaki, and Sabrina Kirrane. GUCON: A Generic Graph Pattern Based Policy Framework for Usage Control Enforcement. In Anna Fensel, Ana Ozaki, Dumitru Roman, and Ahmet Soylu, editors, *Rules and Reasoning*, Lecture Notes in Computer Science, pages 34–53. Springer Nature Switzerland, 2023. ISBN 978-3-031-45072-3. doi: 10.1007/978-3-031-45072-3_3.

Riccardo Albertoni, David Browning, Simon Cox, Alejandra Gonzalez Beltran, Andrea Perego, and Peter Winstanley. Data Catalog Vocabulary (DCAT) Version 2 – W3C Recommendation 04 February 2020, 2020. URL https://www.w3.org/TR/vocab-dcat-2/.

George Alter, Alejandra Gonzalez-Beltran, Lucila Ohno-Machado, and Philippe Rocca-Serra. The Data Tags Suite (DATS) model for discovering data access and use requirements. *GigaScience*, 9 (2), 2020. ISSN 2047-217X. doi: 10.1093/gigascience/giz165.

Muhammad Amith, Marcelline R Harris, Cooper Stansbury, Kathleen Ford, Frank J Manion, and Cui Tao. Expressing and Executing Informed Consent Permissions Using SWRL: The All of Us

Use Case. *AMIA Annual Symposium Proceedings*, pages 197–206, 2022. ISSN 1942-597X. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8861693/.

Nariman Ammar, James E. Bailey, Robert L. Davis, and Arash Shaban-Nejad. The Personal Health Library: A Single Point of Secure Access to Patient Digital Health Information. In *Digital Personalized Health and Medicine*, pages 448–452. IOS Press, 2020. doi: 10.3233/SHTI200200.

Nariman Ammar, James E Bailey, Robert L Davis, and Arash Shaban-Nejad. Using a Personal Health Library–Enabled mHealth Recommender System for Self-Management of Diabetes Among Underserved Populations: Use Case for Knowledge Graphs and Linked Data. *JMIR Formative Research*, 5(3), 2021. ISSN 2561-326X. doi: 10.2196/24738.

Julio Angulo, Simone Fischer-Hübner, Erik Wästlund, and Tobias Pulls. Towards usable privacy policy display and management. *Information Management & Computer Security*, 20(1):4–17, 2012. ISSN 0968-5227. doi: 10.1108/09685221211219155.

Alberto Apollaro, Sophie Aubin, Cristina Azorín, Paola Azrilevich, Isabel Bernal, Dan Liu, Dom Fripp, Gültekin Gürdal, Sawsan Habre, Juha Hakala, Yutaka Hayashi, Ilkay Holt, Nie Hua, Hilary Jones, Tomoko Kataoka, Laurence Le Borgne, Ku (Alan) Liping, Marina Losada, Susanna Mornati, Brigit Nonó, Milan Ojsteršek, Pedro Príncipe, Jochen Schirrwagen, Milica Sevkusic, Tomoya Shiota, Iryna Solodovnik, Wilko Steinhoff, and Nathalie Vedovotto. Controlled Vocabularies for Repositories: Access Rights 1.1, 2022. URL http://purl.org/coar/access_right/.

Claudio A Ardagna, Laurent Bussard, Sabrina De Capitani di Vimercati, Gregory Neven, Stefano Paraboschi, Eros Pedrini, Franz-Stefan Preiss, Dave Raggett, Pierangela Samarati, Slim Trabelsi, and Mario Verdicchio. PrimeLife Policy Language. Technical report, 2009.

Dörthe Arndt, Jeen Broekstra, Bob DuCharme, Ora Lassila, Peter F. Patel-Schneider, Eric Prud'hommeaux, Ted Thibodeau, Jr., and Bryan Thompson. RDF-star and SPARQL-star. *Draft Community Group Report*, 2023. URL https://w3c.github.io/rdf-star/cg-spec/editors_draft.html.

Article 29 Data Protection Working Party. Working Document on the processing of personal data relating to health in electronic health records (EHR), 2007. URL https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2007/wp131_en.pdf.

Article 29 Data Protection Working Party. Opinion 15/2011 on the definition of consent, 2011. URL https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2011/wp187_en.pdf.

Article 29 Data Protection Working Party. Opinion 03/2013 on purpose limitation, 2013. URL https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2013/wp203_en.pdf.

Article 29 Data Protection Working Party. Article 29 Data Protection Working Party comments in response to W3C's public consultation on the W3C Last Call Working Draft, 24 April 2014, Tracking Preference Expression (DNT), 2014. URL https://ec.europa.eu/justice/article-29/documentation/other-

document/files/2014/20140606_wp29_ts_standardisation_letter_to_w3c.pdf.

Article 29 Data Protection Working Party. Guidelines on consent under Regulation 2016/679, 2016. URL https://ec.europa.eu/newsroom/article29/redirection/document/51030.

Article 29 Data Protection Working Party. Guidelines on transparency under Regulation 2016/679, 2018. URL https://ec.europa.eu/newsroom/article29/items/622227.

Haleh Asgarinia, Andres Chomczyk Penedo, Beatriz Esteves, and Dave Lewis. "Who Should I Trust with My Data?" Ethical and Legal Challenges for Innovation in New Decentralized Data Management Technologies. *Information*, 14(7), 2023. ISSN 2078-2489. doi: 10.3390/info14070351.

Paul Ashley, Satoshi Hada, Günter Karjoth, and Matthias Schunter. E-P3P privacy policies and privacy authorization. In *Proceeding of the ACM workshop on Privacy in the Electronic Society - WPES '02*, pages 103–109. ACM Press, 2002. ISBN 978-1-58113-633-3. doi: 10.1145/644527.644538.

Paul Ashley, Satoshi Hada, Günter Karjoth, Calvin Powers, and Matthias Schunter. Enterprise Privacy Authorization Language (EPAL 1.2), 2003. URL https://www.w3.org/Submission/2003/SUBM-EPAL-20031110/.

Jef Ausloos, Rene Mahieu, and Michael Veale. Getting Data Subject Rights Right. *Journal of Intellectual Property, Information Technology and Electronic Commerce Law*, 10:283, 2019.

Monir Azraoui, Kaoutar Elkhiyaoui, Melek Önen, Karin Bernsmed, Anderson Santana De Oliveira, and Jakub Sendor. A-PPL: An Accountability Policy Language. Research report, 2014. URL http://www.eurecom.fr/en/publication/4372/download/rs-publi-4372.pdf.

Tim Baarslag, Alper T Alan, Richard Gomer, and Muddasser Alam. An Automated Negotiation Agent for Permission Management. volume 1, pages 380–390. International Foundation for Autonomous Agents and Multiagent Systems (IFAAMAS), 2017.

Hadrien Bailly, Anoop Papanna, and Rob Brennan. Prototyping an End-User User Interface for the Solid Application Interoperability Specification Under GDPR. In Catia Pesquita, Ernesto Jimenez-Ruiz, Jamie McCusker, Daniel Faria, Mauro Dragoni, Anastasia Dimou, Raphael Troncy, and Sven Hertling, editors, *The Semantic Web*, Lecture Notes in Computer Science, pages 557–573. Springer Nature Switzerland, 2023. ISBN 978-3-031-33455-9. doi: 10.1007/978-3-031-33455-9_33.

Virginia Balseiro, Timea Turdean, and Jeff Zucker. Solid WebID Profile Version 1.0.0. *W3C Community Group Draft Report*, 2022. URL https://solid.github.io/webid-profile/.

Gioele Barabucci, Luca Cervone, Angelo Di Iorio, Monica Palmirani, Silvio Peroni, and Fabio Vitali. Managing semantics in XML vocabularies: an experience in the legal and legislative domain. In *Balisage: The Markup Conference 2010*, volume 5, 2010. ISBN 978-1-935958-01-7. doi: 10.4242/BalisageVol5.Barabucci01.

Cesare Bartolini and Robert Muthuri. Reconciling Data Protection Rights and Obligations: An Ontology of the Forthcoming EU Regulation. In *Workshop on Language and Semantic Technology for Legal Domain*, 2015.

Cesare Bartolini, Robert Muthuri, and Cristiana Santos. Using Ontologies to Model Data Protection Requirements in Workflows. In Mihoko Otake, Setsuya Kurahashi, Yuiko Ota, Ken Satoh, and Daisuke Bekki, editors, *New Frontiers in Artificial Intelligence*, volume 10091 of *Lecture Notes in Computer Science*, pages 233–248. Springer International Publishing, 2017. ISBN 978-3-319-50952-5 978-3-319-50953-2. doi: 10.1007/978-3-319-50953-2_17.

Moritz Becker, Cedric Fournet, and Andrew Gordon. Design and Semantics of a Decentralized Authorization Language. In *20th IEEE Computer Security Foundations Symposium (CSF'07)*, pages 3–15. IEEE, 2007. ISBN 978-0-7695-2819-9. doi: 10.1109/CSF.2007.18. ISSN: 1063-6900.

Moritz Y Becker, Alexander Malkis, and Laurent Bussard. A Framework for Privacy Preferences and Data-Handling Policies. Technical report, Microsoft Research, 2009. URL https://www.microsoft.com/en-us/research/wp-content/uploads/2009/09/A-Framework-for-Privacy-Preferences-and-Data-Handling-Policies-2009-09-28.pdf.

Moritz Y. Becker, Alexander Malkis, and Laurent Bussard. S4P: A Generic Language for Specifying Privacy Preferences and Policies. Technical report, Microsoft Research, 2010. URL https://www.microsoft.com/en-us/research/wp-content/uploads/2010/04/main-1.pdf.

Omri Ben-Shahar and Carl E. Schneider. *More Than You Wanted to Know: The Failure of Mandated Disclosure*. Princeton University Press, 2014. ISBN 978-1-4008-5038-9. doi: 10.1515/9781400850389.

Anders Berglund, Scott Boag, Don Chamberlin, Mary F. Fernández, Michael Kay, Jonathan Robie, and Jérôme Siméon. XML Path Language (XPath) 2.0 (Second Edition), 2010. URL https://www.w3.org/TR/xpath20/.

Tim Berners-Lee, James Hendler, and Ora Lassila. The Semantic Web. *Scientific American*, 2001. URL http://web.archive.org/web/20070713230811/http://www.sciam.com/print_version.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21.

Tim Berners-Lee, Dan Connolly, Lalana Kagal, Yosi Scharf, and Jim Hendler. N3Logic: A logical framework for the World Wide Web. In *Theory and Practice of Logic Programming*, volume 8, pages 249–269, 2008. doi: 10.1017/S1471068407003213.

Stefan Berthold. Towards a Formal Language for Privacy Options. In S. Fischer-Hübner, P. Duquenoy, M. Hansen, R. Leenes, and G. Zhang, editors, *Privacy and Identity 2010: Privacy and Identity Management for Life*, volume 352 of *IFIP Advances in Information and Communication Technology*, pages 27–40. Springer, Berlin, Heidelberg, 2011. URL https://link.springer.com/chapter/10.1007/978-3-642-20769-3_3.

Stefan Berthold. The Privacy Option Language - Specification & Implementation. Research report, Faculty of Health, Science and Technology, Karlstad University, 2013. URL http://kau.diva-portal.org/smash/get/diva2:623452/FULLTEXT01.pdf.

Justin Bingham, Eric Prud'hommeaux, and elf Pavlik. Solid Application Interoperability.

W3C Community Group Draft Report, 2023. URL https://solid.github.io/data-interoperability-panel/specification/.

Balázs Bodó. Mediated trust: A theoretical framework to address the trustworthiness of technological trust mediators. *New Media & Society*, 23(9):2668–2690, 2021. ISSN 1461-4448. doi: 10.1177/1461444820939922.

Sarah N. Boers, Johannes J. M. van Delden, and Annelien L. Bredenoord. Broad Consent Is Consent for Governance. *The American Journal of Bioethics*, 15(9):53–55, 2015. ISSN 1526-5161. URL https://doi.org/10.1080/15265161.2015.1062165.

Kathy Bohrer and Bobby Holland. Customer Profile Exchange (CPExchange) Specification. Technical specification, 2000.

Harold Boley, Adrian Paschke, Tara Athan, Adrian Giurca, Nick Bassiliades, Guido Governatori, Monica Palmirani, Adam Wyner, Alexander Kozlenkov, and Gen Zou. Specification of RuleML 1.02, 2017. URL http://wiki.ruleml.org/index.php/Specification_of_RuleML_1.02.

P.A. Bonatti, S. Kirrane, I. Petrova, L. Sauro, and E. Schlehahn. Policy Language V2 – Deliverable D2.5. Project deliverable, 2018a. URL https://www.specialprivacy.eu/images/documents/SPECIAL_D25_M21_V10.pdf.

Piero Bonatti, Sabrina Kirrane, Iliana Mineva Petrova, Luigi Sauro, and Eva Schlehahn. The SPECIAL Usage Policy Language version 1.0, 2019. URL https://ai.wu.ac.at/policies/policylanguage/.

Piero A Bonatti, Bert Bos, Stefan Decker, Javier D Fernandez, Sabrina Kirrane, Vassilios Peristeras, Axel Polleres, and Rigo Wenning. Data Privacy Vocabularies and Controls: Semantic Web for Transparency and Privacy. In *Semantic Web for Social Good (SWSG2018) @ ISWC2018*. CEUR Workshop Proceedings, 2018b.

Piero A. Bonatti, Luca Ioffredo, Iliana M. Petrova, Luigi Sauro, and Ida R. Siahaan. Real-time reasoning in OWL2 for GDPR compliance. *Artificial Intelligence*, 289, 2020. ISSN 0004-3702. doi: 10.1016/j.artint.2020.103389.

Matthieu Bosquet. Access Control Policy (ACP) Version 0.9.0. *W3C Community Group Draft Report*, 2022. URL https://solidproject.org/TR/acp.

Anu Bradford. *The Brussels Effect: How the European Union Rules the World*. Oxford University Press, 2019. ISBN 978-0-19-008858-3. doi: 10.1093/oso/9780190088583.003.0002.

Christoph H.-J. Braun and Tobias Käfer. Attribute-based Access Control on Solid Pods using Privacy-friendly Credentials. In *Proceedings of the Poster and Demo Track and Workshop Track of the 18th International Conference on Semantic Systems Co-Located with 18th International Conference on Semantic Systems (SEMANTiCS 2022)*, 2022a.

Christoph H.-J. Braun and Tobias Käfer. Self-verifying Web Resource Representations Using Solid, RDF-Star and Signed URIs. In Paul Groth, Anisa Rula, Jodi Schneider, Ilaria Tiddi, Elena Simperl, Panos Alexopoulos, Rinke Hoekstra, Mehwish Alam, Anastasia Dimou, and Minna Tamper, editors, *The Semantic Web: ESWC 2022 Satellite Events*, Lecture Notes in Computer

Science, pages 138–142. Springer International Publishing, 2022b. ISBN 978-3-031-11609-4. doi: 10.1007/978-3-031-11609-4_26.

Dan Brickley and R.V. Guha. RDF Schema 1.1. *W3C Recommendation*, 2014. URL https://www.w3.org/TR/rdf11-schema/.

Dan Brickley and Libby Miller. FOAF Vocabulary Specification, 2004. URL http://xmlns.com/foaf/0.1/.

Simon Brown. *The C4 model for visualising software architecture*. Leanpub, 2015. URL https://leanpub.next/visualising-software-architecture.

Raf Buyle, Ruben Taelman, Katrien Mostaert, Geroen Joris, Erik Mannens, Ruben Verborgh, and Tim Berners-Lee. Streamlining Governmental Processes by Putting Citizens in Control of Their Personal Data. In Andrei Chugunov, Igor Khodachek, Yuri Misnikov, and Dmitrii Trutnev, editors, *Electronic Governance and Open Society: Challenges in Eurasia*, volume 1135, pages 346–359. Springer International Publishing, 2020. ISBN 978-3-030-39295-6 978-3-030-39296-3. doi: 10.1007/978-3-030-39296-3_26. Communications in Computer and Information Science.

John Byrum, Suzanne Jouguelet, Dorothy McGarry, Nancy Williamson, Maria Witt, Tom Delsey, Elizabeth Dulabahn, Elaine Svenonius, and Barbara Tillett. Functional Requirements for Bibliographic Records. Technical report, 2009. URL https://www.ifla.org/publications/functional-requirements-for-bibliographic-records.

Moran N. Cabili, Jonathan Lawson, Andrea Saltzman, Greg Rushton, Pearl O'Rourke, John Wilbanks, Laura Lyman Rodriguez, Tommi Nyronen, Mélanie Courtot, Stacey Donnelly, and Anthony A. Philippakis. Empirical validation of an automated approach to data use oversight. *Cell Genomics*, 1(2), 2021. ISSN 2666-979X. doi: 10.1016/j.xgen.2021.100031.

Juan Cano-Benito, Andrea Cimmino, and Raúl García-Castro. Injecting data into ODRL privacy policies dynamically with RDF mappings. In *Companion Proceedings of the ACM Web Conference 2023*, pages 246–249. Association for Computing Machinery, 2023. ISBN 978-1-4503-9419-2. doi: 10.1145/3543873.3587358. URL https://doi.org/10.1145/3543873.3587358.

Sarven Capadisli. Web Access Control Version 1.0.0. *W3C Candidate Recommendation*, 2022. URL https://solidproject.org/TR/wac.

Sarven Capadisli and Amy Guy. Linked Data Notifications. *W3C Recommendation*, 2017. URL https://www.w3.org/TR/ldn/.

Sarven Capadisli, Tim Berners-Lee, Ruben Verborgh, and Kjetil Kjernsmo. Solid Protocol Version 0.10.0. *W3C Community Group Draft Report*, 2022. URL https://solidproject.org/TR/protocol.

Gabriele Carovano and Michèle Finck. Regulating Data Intermediaries: The Impact of the Data Governance Act on the EU's Data Economy, 2023.

Núria Casellas, Juan-Emilio Nieto, Albert Meroño, Antoni Roig, Sergi Torralba, Mario Reyes, and Pompeu Casanovas. Ontological Semantics for Data Privacy Compliance: The NEURONA Project. In *2010 AAAI Spring Symposium*, Intelligent Information Privacy Management, pages

34–38. AAAI, 2010. URL https://ddd.uab.cat/pub/artpub/2010/137891/aaaisprsymser_a2010n1iENG.pdf.

Fred H. Cate. The Failure of Fair Information Practice Principles. In *Consumer Protection in the Age of the 'Information Economy'*. Routledge, 2006. ISBN 978-1-315-57371-7.

Andrés Chomczyk Penedo. Self-sovereign identity systems and European data protection regulations: an analysis of roles and responsibilities. In *Open Identity Summit 2021*, pages 95–106. Gesellschaft für Informatik e.V., 2021. ISBN 978-3-88579-706-7. URL http://dl.gi.de/handle/20.500.12116/36505.

Andrés Chomczyk Penedo. Towards a technologically assisted consent in the upcoming new EU data laws? *Privacy in Germany*, 5:180–187, 2022. ISSN 2196-9817. doi: 10.37307/j.2196-9817.2022.05.05.

Serge Chávez-Feria, Raúl García-Castro, and María Poveda-Villalón. Chowlk: from UML-Based Ontology Conceptualizations to OWL. In Paul Groth, Maria-Esther Vidal, Fabian Suchanek, Pedro Szekley, Pavan Kapanipathi, Catia Pesquita, Hala Skaf-Molli, and Minna Tamper, editors, *The Semantic Web: 19th International Conference, ESWC 2022*, Lecture Notes in Computer Science, pages 338–352. Springer International Publishing, 2022. ISBN 978-3-031-06981-9. doi: 10.1007/978-3-031-06981-9_20.

Paolo Ciccarese, Stian Soiland-Reyes, Khalid Belhajjame, Alasdair J. G. Gray, Carole Goble, and Tim Clark. PAV ontology: provenance, authoring and versioning. *Journal of Biomedical Semantics*, 4 (37), 2013. ISSN 2041-1480. doi: 10.1186/2041-1480-4-37.

James Clark and Steve DeRose. XML Path Language (XPath) Version 1.0, 1999. URL https://www.w3.org/TR/1999/REC-xpath-19991116/.

Aaron Coburn, elf Pavlik, and Dmitri Zagidulin. Solid-OIDC Version 0.1.0. *W3C Community Group Draft Report*, 2022. URL https://solidproject.org/TR/oidc.

Jessica Colnago, Lorrie Faith Cranor, Alessandro Acquisti, and Kate Hazel Stanton. Is it a concern or a preference? An investigation into the ability of privacy scales to capture and distinguish granular privacy constructs. In *18th Symposium on Usable Privacy and Security (SOUPS 2022)*, pages 331–346, 2022. ISBN 978-1-939133-30-4. URL https://www.usenix.org/conference/soups2022/presentation/colnago.

Council of Europe. European Convention on Human Rights, 1950. URL https://www.echr.coe.int/Documents/Convention_ENG.pdf.

Council of Europe. Convention 108+: Convention for the protection of individuals with regard to the processing of personal data, 1981. URL https://www.coe.int/en/web/data-protection/convention108-and-protocol.

Massimo Craglia, Henk Scholten, Marina Micheli, Jiri Hradec, Igor Calzada, Steven Luitjens, Marisa Ponti, and Jaap Boter. *Digitranscope: The governance of digitally transformed society*. Publications Office of the European Union, 2021. ISBN 978-92-76-30229-2. doi: 10.2760/503546.

Lorrie Cranor. *Web Privacy with P3P*. O'Reilly Media, Inc., 1st edition, 2002. ISBN 978-0-596-00371-5.

Lorrie Cranor, Marc Langheinrich, and Massimo Marchiori. A P3P Preference Exchange Language 1.0 (APPEL1.0), 2002a. URL https://www.w3.org/TR/2002/WD-P3P-preferences-20020415/.

Lorrie Cranor, Marc Langheinrich, Massimo Marchiori, Martin Presler-Marshall, and Joseph Reagle. The Platform for Privacy Preferences 1.0 (P3P1.0) Specification – W3C Recommendation 16 April 2002 obsoleted 30 August 2018, 2002b. URL https://www.w3.org/TR/P3P/.

R. Jason Cronk. Categories of personal information, 2017. URL https://enterprivacy.com/2017/03/01/categories-of-personal-information/.

Richard Cyganiak, David Wood, and Markus Lanthaler. RDF 1.1 Concepts and Abstract Syntax. *W3C Recommendation*, 2014. URL https://www.w3.org/TR/rdf11-concepts/.

Marcos Cáceres, Kenneth Rohde Christiansen, Matt Giuca, Aaron Gustafson, Daniel Murphy, Anssi Kostiainen, Mounir Lamouri, and Rob Dolin. Web Application Manifest – W3C Working Draft 15 November 2023, 2023. URL https://www.w3.org/TR/appmanifest/.

DCMI Usage Board. DCMI Metadata Terms, 2020. URL https://www.dublincore.org/specifications/dublin-core/dcmi-terms/.

D. De Bot and T. Haegemans. Data Sharing Patterns as a Tool to Tackle Legal Considerations about Data Reuse with Solid: Theory and Applications in Europe. Digita research reports, 2021. URL https://go.digita.ai/reuse-patterns.

Primavera De Filippi, Morshed Mannan, and Wessel Reijers. Blockchain as a confidence machine: The problem of trust & challenges of governance. *Technology in Society*, 62, 2020. ISSN 0160-791X. doi: 10.1016/j.techsoc.2020.101284.

Gertjan De Mulder, Ben De Meester, Pieter Heyvaert, Ruben Taelman, Anastasia Dimou, and Ruben Verborgh. PROV4ITDaTa: Transparent and direct transfer of personal data to personal stores. In *Companion Proceedings of the Web Conference 2021*, WWW '21, pages 695–697. Association for Computing Machinery, 2021. ISBN 978-1-4503-8313-4. doi: 10.1145/3442442.3458608.

Marina De Vos, Sabrina Kirrane, Julian Padget, and Ken Satoh. ODRL Policy Modelling and Compliance Checking. In Paul Fodor, Marco Montali, Diego Calvanese, and Dumitru Roman, editors, *Rules and Reasoning*, Lecture Notes in Computer Science, pages 36–51. Springer International Publishing, 2019. ISBN 978-3-030-31095-0. doi: 10.1007/978-3-030-31095-0_3.

Ruben Dedecker, Wout Slabbinck, Jesse Wright, Patrick Hochstenbach, Pieter Colpaert, and Ruben Verborgh. What's in a Pod? A Knowledge Graph Interpretation For The Solid Ecosystem. In *Proceedings of the 6th Workshop on Storing, Querying and Benchmarking Knowledge Graphs*, volume 3279 of *CEUR Workshop Proceedings*, pages 81–96, 2022. ISBN 1613-0073. URL https://ceur-ws.org/Vol-3279/paper6.pdf.

Anastasia Dimou, Miel Vander Sande, Pieter Colpaert, Ruben Verborgh, Erik Mannens, and Rik Van de Walle. RML: A generic language for integrated RDF mappings of heterogeneous data. In *Proceedings of the 7th Workshop on Linked Data on the Web*, volume 1184. CEUR, 2014. URL http://ceur-ws.org/Vol-1184/ldow2014_paper_01.pdf.

Stephanie O. M. Dyke, Anthony A. Philippakis, Jordi Rambla De Argila, Dina N. Paltoo, Erin S. Luetkemeier, Bartha M. Knoppers, Anthony J. Brookes, J. Dylan Spalding, Mark Thompson, Marco Roos, Kym M. Boycott, Michael Brudno, Matthew Hurles, Heidi L. Rehm, Andreas Matern, Marc Fiume, and Stephen T. Sherry. Consent Codes: Upholding Standard Data Use Conditions. *PLOS Genetics*, 12(1), 2016. ISSN 1553-7404. doi: 10.1371/journal.pgen.1005772.

Lavanya Elluri and Karuna Pande Joshi. A Knowledge Representation of Cloud Data Controls for EU GDPR Compliance. In *2018 IEEE World Congress on Services (SERVICES)*, pages 45–46. IEEE, 2018. ISBN 978-1-5386-7374-4. doi: 10.1109/SERVICES.2018.00036.

Lavanya Elluri, Ankur Nagar, and Karuna Pande Joshi. An Integrated Knowledge Graph to Automate GDPR and PCI DSS Compliance. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 1266–1271. IEEE, 2018. ISBN 978-1-5386-5035-6. doi: 10.1109/BigData.2018.8622236.

Christian Esposito, Ross Horne, Livio Robaldo, Bart Buelens, and Elfi Goesaert. Assessing the Solid Protocol in Relation to Security and Privacy Obligations. *Information*, 14(7), 2023. ISSN 2078-2489. doi: 10.3390/info14070411.

Beatriz Esteves. Challenges in the Digital Representation of Privacy Terms. In Víctor Rodríguez-Doncel, Monica Palmirani, Michał Araszkiewicz, Pompeu Casanovas, Ugo Pagallo, and Giovanni Sartor, editors, *AI Approaches to the Complexity of Legal Systems XI-XII*, volume 13048 of *Lecture Notes in Computer Science*, pages 313–327. Springer International Publishing, 2021. ISBN 978-3-030-89811-3. URL https://doi.org/10.1007/978-3-030-89811-3_22.

Beatriz Esteves. Towards an Architecture for Data Altruism in Solid. In *To Appear on the Proceedings of the 22nd International Semantic Web Conference: Posters, Demos, and Industry Tracks*, 2023.

Beatriz Esteves and Harshvardhan J. Pandit. Using Patterns to Manage Governance of Solid Apps. In *To Appear on the Proceedings of the 14th Workshop on Ontology Design and Patterns (WOP 2023@ISWC 2023)*, 2023. doi: 10.5281/zenodo.8303102. URL https://zenodo.org/record/8303102.

Beatriz Esteves and Víctor Rodríguez-Doncel. Analysis of Ontologies and Policy Languages to Represent Information Flows in GDPR. *Semantic Web*, 2022a.

Beatriz Esteves and Víctor Rodríguez-Doncel. Semantifying the Governance of Data in Europe. In *18th International Conference on Semantic Systems – CEUR Workshop Proceedings*, volume 3235, 2022b. URL https://ceur-ws.org/Vol-3235/paper17.pdf.

Beatriz Esteves, Harshvardhan J. Pandit, and Víctor Rodríguez-Doncel. ODRL Profile for Expressing Consent through Granular Access Control Policies in Solid. In *2021 IEEE European Symposium on Security and Privacy Workshops (EuroS PW)*, pages 298–306, 2021. doi: 10.1109/EuroSPW54576.2021.00038.

Beatriz Esteves, Haleh Asgarinia, Andres Chomczyk Penedo, Blessing Mutiro, and Dave Lewis. Fostering trust with transparency in the data economy era: an integrated ethical, legal, and knowledge engineering approach. In *Proceedings of the 1st International Workshop on Data Economy*, pages 57–63. Association for Computing Machinery, 2022a. ISBN 978-1-4503-9923-4. doi: 10.1145/3565011.3569061.

Beatriz Esteves, Kieran Fraser, Shridhar Kulkarni, Owen Conlan, and Víctor Rodríguez-Doncel. Extracting and Understanding Call-to-actions of Push-Notifications. In Paolo Rosso, Valerio Basile, Raquel Martínez, Elisabeth Métais, and Farid Meziane, editors, *Natural Language Processing and Information Systems*, volume 13286 of *Lecture Notes in Computer Science*, pages 147–159. Springer International Publishing, 2022b. ISBN 978-3-031-08473-7. doi: 10.1007/978-3-031-08473-7_14.

Beatriz Esteves, Kieran Fraser, Shridhar Kulkarni, Owen Conlan, and Víctor Rodríguez-Doncel. Now, Later, Never: A Study of Urgency in Mobile Push-Notifications. In Pari Delir Haghighi, Ismail Khalil, and Gabriele Kotsis, editors, *Advances in Mobile Computing and Multimedia Intelligence*, Lecture Notes in Computer Science, pages 38–44. Springer Nature Switzerland, 2022c. ISBN 978-3-031-20436-4. doi: 10.1007/978-3-031-20436-4_4.

Beatriz Esteves, Victor Rodriguez-Doncel, and Ricardo Longares. Automating the Response to GDPR's Right of Access. In *Legal Knowledge and Information Systems*, pages 170–175. IOS Press, 2022d. doi: 10.3233/FAIA220462. URL https://ebooks.iospress.nl/doi/10.3233/FAIA220462.

Beatriz Esteves, Víctor Rodríguez-Doncel, Harshvardhan J. Pandit, Nicolas Mondada, and Pat McBennett. Using the ODRL Profile for Access Control for Solid Pod Resource Governance. In Paul Groth, Anisa Rula, Jodi Schneider, Ilaria Tiddi, Elena Simperl, Panos Alexopoulos, Rinke Hoekstra, Mehwish Alam, Anastasia Dimou, and Minna Tamper, editors, *The Semantic Web: ESWC 2022 Satellite Events*, Lecture Notes in Computer Science, pages 16–20. Springer International Publishing, 2022e. ISBN 978-3-031-11609-4. doi: 10.1007/978-3-031-11609-4_3.

Beatriz Esteves, Víctor Rodríguez-Doncel, Harshvardhan J. Pandit, and Dave Lewis. Semantics for Implementing Data Reuse and Altruism Under EU's Data Governance Act. In *Knowledge Graphs: Semantics, Machine Learning, and Languages*, pages 210–226. IOS Press, 2023. doi: 10.3233/SSW230015.

European Commission. Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions – A European strategy for data. *COM(2020) 66 final*, 2020. URL https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020DC0066.

European Data Protection Board. Guidelines 2/2019 on the processing of personal data under Article 6(1)(b) GDPR in the context of the provision of online services to data subjects, 2019. URL https://edpb.europa.eu/our-work-tools/our-documents/guidelines/guidelines-22019-processing-personal-data-under-article-61b_en.

European Data Protection Board. Guidelines 05/2020 on consent under Regulation 2016/679 Version 1.1, 2020a. URL https://edpb.europa.eu/sites/edpb/files/files/file1/edpb_guidelines_202005_consent_en.pdf.

European Data Protection Board. Guidelines 07/2020 on the concepts of controller and processor in the GDPR, 2020b. URL https://edpb.europa.eu/sites/default/files/consultation/edpb_guidelines_202007_controllerprocessor_en.pdf.

European Data Protection Supervisor. A Preliminary Opinion on data protection and scientific research, 2020. URL https://edps.europa.eu/sites/edp/files/publication/20-01-06_opinion_research_en.pdf.

European Data Protection Supervisor. TechDispatch #3/2020 - Personal Information Management Systems. Technical report, 2021. URL https://edps.europa.eu/data-protection/our-work/publications/techdispatch/techdispatch-32020-personal-information_en.

European Union Agency for Fundamental Rights and Council of Europe. *Handbook on European data protection law*. Publications Office of the European Union, 2018. ISBN 978-92-871-9849-5 978-92-9491-903-8 978-92-9491-901-4. doi: 10.2811/58814.

Kayode Yadilichi Ezike. SolidVC: a decentralized framework for Verifiable Credentials on the web. *Master's thesis*, 2019. URL https://hdl.handle.net/1721.1/121667. Massachusetts Institute of Technology.

Benjamin Fabian, Tatiana Ermakova, and Tino Lentz. Large-scale readability analysis of privacy policies. In *Proceedings of the International Conference on Web Intelligence*, WI '17, pages 18–25, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 978-1-4503-4951-2. doi: 10.1145/3106426.3106427.

Khalid U. Fallatah, Mahmoud Barhamgi, and Charith Perera. Personal Data Stores (PDS): A Review. *Sensors*, 23(3):1477, 2023. ISSN 1424-8220. doi: 10.3390/s23031477.

Kaniz Fatema, Ensar Hadziselimovic, Harshvardhan Pandit, Christophe Debruyne, Dave Lewis, and Declan O'Sullivan. Compliance through Informed Consent: Semantic Based Consent Permission and Data Management Model. In *Proceedings of the 5th Workshop on Society, Privacy and the Semantic Web - Policy and Technology (PrivOn2017) co-located with the 16th International Semantic Web Conference (ISWC 2017)*, volume 1951. CEUR, 2017. URL https://ceur-ws.org/Vol-1951/PrivOn2017_paper_5.pdf.

Heike Felzmann, Eduard Fosch Villaronga, Christoph Lutz, and Aurelia Tamò-Larrieux. Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns. *Big Data & Society*, 6(1), 2019. ISSN 2053-9517. doi: 10.1177/2053951719860542.

Oliver Fich. Finland changes cookie rules, 2021. URL https://cookieinformation.com/resources/blog/finland-changes-cookie-rules/.

Roy T. Fielding, Mark Nottingham, and Julian Reschke. HTTP Semantics. Internet Standard RFC 9110, Internet Engineering Task Force, 2022. URL https://datatracker.ietf.org/doc/rfc9110.

Marcu Florea and Beatriz Esteves. Is Automated Consent in Solid GDPR-Compliant? An Approach for Obtaining Valid Consent with the Solid Protocol. *Information*, 14(12), 2023. ISSN 2078-2489. doi: 10.3390/info14120631.

Nicoletta Fornara and Marco Colombetti. Operational Semantics of an Extension of ODRL Able to Express Obligations. In Francesco Belardinelli and Estefanía Argente, editors, *Multi-Agent*

Systems and Agreement Technologies, Lecture Notes in Computer Science, pages 172–186. Springer International Publishing, 2018. ISBN 978-3-030-01713-2. doi: 10.1007/978-3-030-01713-2_13.

Nicoletta Fornara and Marco Colombetti. Using Semantic Web technologies and production rules for reasoning on obligations, permissions, and prohibitions. *AI Communications*, 32(4):319–334, 2019. ISSN 0921-7126. doi: 10.3233/AIC-190617.

Nicoletta Fornara, Víctor Rodríguez-Doncel, and Beatriz Esteves. ODRL Formal Semantics – Draft Community Group Report 28 July 2023, 2023. URL https://w3c.github.io/odrl/formal-semantics/.

Aldo Gangemi, Silvio Peroni, David Shotton, and Fabio Vitali. The Publishing Workflow Ontology (PWO). *Semantic Web*, 8(5):703–718, 2017. ISSN 1570-0844. doi: 10.3233/SW-160230. Publisher: IOS Press.

Daniel Garijo and Yolanda Gil. Augmenting PROV with Plans in P-PLAN: Scientific Processes as Linked Data. In *CEUR Workshop Proceedings*, 2012.

Daniel Garijo, Oscar Corcho, and Maria Poveda-Villalon. FOOPS!: An Ontology Pitfall Scanner for the FAIR principles. In *International Semantic Web Conference (ISWC) 2021: Posters, Demos, and Industry Tracks*, volume 2980. CEUR Workshop Proceedings, 2021. URL http://ceur-ws.org/Vol-2980/paper321.pdf.

Armin Gerl. Extending Layered Privacy Language to Support Privacy Icons for a Personal Privacy Policy User Interface. 2018. doi: 10.14236/ewic/hci2018.177.

Armin Gerl and Bianca Meier. Privacy in the Future of Integrated Health Care Services – Are Privacy Languages the Key? In *2019 International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*, pages 312–317, 2019. doi: 10.1109/WiMOB.2019.8923532.

Armin Gerl and Dirk Pohl. Critical Analysis of LPL according to Articles 12 - 14 of the GDPR. pages 1–9, 2018. doi: 10.1145/3230833.3233267.

Armin Gerl and Florian Prey. LPL Personal Privacy Policy User Interface: Design and Evaluation. 2018. doi: 10.18420/MUC2018-WS08-0540. Publisher: Gesellschaft für Informatik e.V.

Armin Gerl, Nadia Bennani, Harald Kosch, and Lionel Brunie. LPL, Towards a GDPR-Compliant Privacy Language: Formal Definition and Usage. In Abdelkader Hameurlain and Roland Wagner, editors, *Transactions on Large-Scale Data- and Knowledge-Centered Systems XXXVII*, Lecture Notes in Computer Science, pages 41–80. Springer, 2018. ISBN 978-3-662-57932-9. doi: 10.1007/978-3-662-57932-9_2.

Armin Gerl, Bianca Meier, and Stefan Becher. Let Users Control Their Data – Privacy Policy-Based User Interface Design. In Tareq Ahram, Redha Taiar, Serge Colson, and Arnaud Choplin, editors, *Human Interaction and Emerging Technologies*, pages 790–795, Cham, 2020. Springer International Publishing. ISBN 978-3-030-25629-6. doi: 10.1007/978-3-030-25629-6_123.

Soledad Gesteira. Más allá de la apropiación criminal de niños: el surgimiento de organizaciones de personas "adoptadas" que buscan su "identidad biológica" en Argentina. *RUNA, archivo para las ciencias del hombre*, 35(1):61–76, 2014. ISSN 1851-9628. doi: 10.34096/runa.v35i1.604.

Alexandra Giannopoulou. Digital Identity Infrastructures: a Critical Approach of Self-Sovereign Identity. *Digital Society*, 2(18), 2023. ISSN 2731-4669. doi: 10.1007/s44206-023-00049-z.

Susan E. Gindin. Nobody Reads Your Privacy Policy or Online Contract? Lessons Learned and Questions Raised by the FTC's Action against Sears. *Northwestern Journal of Technology and Intellectual Property*, 8(1), 2009.

Global Privacy Control. GPC Privacy Browser Signal Now Used by Millions and Honored By Major Publishers, 2021. URL https://globalprivacycontrol.org/press-release/20210128.

Sam Grabus and Jane Greenberg. The Landscape of Rights and Licensing Initiatives for Data Sharing. *Data Science Journal*, 18(1), 2019. ISSN 1683-1470. doi: 10.5334/dsj-2019-029.

Gregg Kellogg, Pierre-Antoine Champin, and Dave Longley. JSON-LD 1.1: A JSON-based Serialization for Linked Data. *W3C Recommendation*, 2020. URL https://www.w3.org/TR/json-ld/.

Morane Gruenpeter, Sabrina Granger, Alain Monteil, Neil Chue Hong, Elena Breitmoser, Mario Antonioletti, Daniel Garijo, Esteban González Guardia, Alejandra Gonzalez Beltran, Carole Goble, Stian Soiland-Reyes, Nick Juty, and Gabriela Mejias. D4.4 – Guidelines for recommended metadata standard for research software within EOSC. Technical report, 2024. URL https://zenodo.org/records/10786147.

R.V. Guha, Dan Brickley, and Steve Macbeth. Schema.org: Evolution of Structured Data on the Web. *ACM Queue*, 13(9), 2015. URL https://queue.acm.org/detail.cfm?id=2857276.

Aaron Gustafson. Web App Manifest – Application Information – W3C Group Note 21 August 2023, 2023. URL https://www.w3.org/TR/manifest-app-info/.

Matilda A. Haas, Harriet Teare, Megan Prictor, Gabi Ceregra, Miranda E. Vidgen, David Bunker, Jane Kaye, and Tiffany Boughtwood. 'CTRL': an online, Dynamic Consent and participant engagement platform working towards solving the complexities of consent in genomic research. *European Journal of Human Genetics*, 29(4):687–698, 2021. ISSN 1476-5438. doi: 10.1038/s41431-020-00782-w.

Steve Harris and Andy Seaborne. SPARQL 1.1 Query Language. *W3C Recommendation*, 2013. URL https://www.w3.org/TR/sparql11-query/.

Giray Havur, Miel Sande, and Sabrina Kirrane. Greater Control and Transparency in Personal Data Processing. In *Proceedings of the 6th International Conference on Information Systems Security and Privacy*, pages 655–662. SCITEPRESS - Science and Technology Publications, 2020. ISBN 978-989-758-399-5. doi: 10.5220/0009143206550662.

Katherine Hawley. *How To Be Trustworthy*. Oxford University Press, 2019. URL https://academic-oup-com.ezproxy2.utwente.nl/book/32233.

Patrick Hochstenbach, Jos De Roo, and Ruben Verborgh. RDF Surfaces: Computer Says No. In *1st Workshop on Trusting Decentralised Knowledge Graphs and Web Data*, 2023. URL https://ruben.verborgh.org/publications/hochstenbach_trusdekw_2023/.

Rinke Hoekstra, Joost Breuker, Marcello Di Bello, and Alexander Boer. The LKIF Core Ontology of Basic Legal Concepts. In Pompeu Casanovas, Maria Angela Biasiotti, Enrico Francesconi, and Maria Teresa Sagri, editors, *Proceedings of the Workshop on Legal Ontologies and Artificial Intelligence Techniques (LOAIT 2007)*, pages 43–63, 2007.

Soheil Human, Max Schrems, Alan Toner, Gerben, and Ben Wagner. Advanced Data Protection Control (ADPC), 2021. URL https://www.dataprotectioncontrol.org/spec/.

R. Iannella, M. Steidl, S. Myles, and V. Rodríguez-Doncel. ODRL Vocabulary & Expression 2.2 – W3C Recommendation 15 February 2018, 2018. URL https://www.w3.org/TR/odrl-vocab/.

Renato Iannella and Serena Villata. ODRL Information Model 2.2 – W3C Recommendation 15 February 2018, 2018. URL https://www.w3.org/TR/odrl-model/.

Luukas K. Ilves and David Osimo. A roadmap for a fair data economy. Policy Brief, Sitra and the Lisbon Council, 2019.

ISO/IEC JTC 1/SC 27. ISO/IEC TS 27560 – Privacy technologies – Consent record information structure, 2023. URL https://www.iso.org/standard/80392.html.

Johnson Iyilade and Julita Vassileva. A Framework for Privacy-Aware User Data Trading. In Sandra Carberry, Stephan Weibelzahl, Alessandro Micarelli, and Giovanni Semeraro, editors, *User Modeling, Adaptation, and Personalization*, volume 7899 of *Lecture Notes in Computer Science*, pages 310–317. Springer Berlin Heidelberg, 2013. ISBN 978-3-642-38843-9 978-3-642-38844-6. doi: 10.1007/978-3-642-38844-6_28.

Johnson Iyilade and Julita Vassileva. P2U: A Privacy Policy Specification Language for Secondary Data Sharing and Usage. In *2014 IEEE Security and Privacy Workshops*, pages 18–22. IEEE, 2014. ISBN 978-1-4799-5103-1. doi: 10.1109/SPW.2014.12.

Michael G. Jacobides, Arun Sundararajan, and Marshall Van Alstyne. *Platforms and Ecosystems: Enabling the Digital Economy*. World Economic Forum, Switzerland, 2019. URL https://www3.weforum.org/docs/WEF_Digital_Platforms_and_Ecosystems_2019.pdf.

Vikas Jaiman and Visara Urovi. A Consent Model for Blockchain-Based Health Data Sharing Platforms. *IEEE Access*, 8:143734–143745, 2020. ISSN 2169-3536. doi: 10.1109/ACCESS.2020.3014565.

Heleen Janssen, Jennifer Cobbe, and Jatinder Singh. Personal information management systems: a user-centric privacy utopia? *Internet Policy Review*, 9(4), 2020. ISSN 2197-6775. URL https://policyreview.info/articles/analysis/personal-information-management-systems-user-centric-privacy-utopia.

Luiza Jarovsky. Improving Consent in Information Privacy Through Autonomy-Preserving Protective Measures (APPMs). *European Data Protection Law Review*, 4(4):447–458, 2018. URL https://papers.ssrn.com/abstract=3350052.

Irene Kamara and Eleni Kosta. Do Not Track initiatives: regaining the lost user control. *International Data Privacy Law*, 6(4):276–290, 2016. ISSN 2044-3994. doi: 10.1093/idpl/ipw019.

Milen G. Kebede, Giovanni Sileno, and Tom Van Engers. A Critical Reflection on ODRL. In Víctor Rodríguez-Doncel, Monica Palmirani, Michał Araszkiewicz, Pompeu Casanovas, Ugo Pagallo, and Giovanni Sartor, editors, *AI Approaches to the Complexity of Legal Systems XI-XII*, Lecture Notes in Computer Science, pages 48–61, Cham, 2021. Springer International Publishing. ISBN 978-3-030-89811-3. doi: 10.1007/978-3-030-89811-3_4.

Ankesh Khandelwal, Jie Bao, Lalana Kagal, Ian Jacobi, Li Ding, and James Hendler. Analyzing the AIR Language: A Semantic Web (Production) Rule Language. In Pascal Hitzler and Thomas Lukasiewicz, editors, *Web Reasoning and Rule Systems*, volume 6333 of *Lecture Notes in Computer Science*, pages 58–72. Springer Berlin Heidelberg, 2010. ISBN 978-3-642-15917-6 978-3-642-15918-3. doi: 10.1007/978-3-642-15918-3_6.

Sabrina Kirrane, Alessandra Mileo, and Stefan Decker. Access Control and the Resource Description Framework: A Survey. *Semantic Web*, 8(2):311–352, 2017. ISSN 1570-0844. doi: 10.3233/SW-160236.

Sabrina Kirrane, Javier D. Fernández, Wouter Dullaert, Uros Milosevic, Axel Polleres, Piero A. Bonatti, Rigo Wenning, Olha Drozd, and Philip Raschke. A Scalable Consent, Transparency and Compliance Architecture. In Aldo Gangemi, Anna Lisa Gentile, Andrea Giovanni Nuzzolese, Sebastian Rudolph, Maria Maleshkova, Heiko Paulheim, Jeff Z Pan, and Mehwish Alam, editors, *The Semantic Web: ESWC 2018 Satellite Events*, volume 11155 of *Lecture Notes in Computer Science*, pages 131–136. Springer International Publishing, 2018a. ISBN 978-3-319-98191-8 978-3-319-98192-5. doi: 10.1007/978-3-319-98192-5_25.

Sabrina Kirrane, Uros Milosevic, Javier D. Fernández, Axel Polleres, and Jonathan Langens. Transparency Framework V2 – Deliverable D2.7. Project deliverable, 2018b. URL https://www.specialprivacy.eu/images/documents/SPECIAL_D27_M23_V10.pdf.

Barbara Kitchenham and Pearl Brereton. A systematic review of systematic review process research in software engineering. *Information and Software Technology*, 55(12):2049–2075, 2013. ISSN 0950-5849. doi: 10.1016/j.infsof.2013.07.010.

Holger Knublauch and Dimitris Kontokostas. Shapes Constraint Language (SHACL) – W3C Recommendation 20 July 2017, 2017. URL https://www.w3.org/TR/shacl/.

Merel Koning. The Purpose and Limitation of the Purpose Limitation Principle. *Doctoral Thesis. Radboud University Nijmegen*, 2020. URL https://repository.ubn.ru.nl/bitstream/handle/2066/221665/221665.pdf?sequence=1.

Bert-Jaap Koops. The concept of function creep. *Law, Innovation and Technology*, 13(1):29–56, 2021. ISSN 1757-9961. doi: 10.1080/17579961.2021.1898299.

Eleni Kosta. *Consent in European Data Protection Law*. Martinus Nijhoff Publishers, 2013. ISBN 978-90-04-23236-5.

Herke R. Kranenborg. Article 8 – Protection of Personal Data. In *The EU Charter of Fundamental Rights*. Hart Publishing, 2014. URL https://www.bloomsbury.com/uk/eu-charter-of-fundamental-rights-9781509933501/.

Katsiaryna Krasnashchok, Majd Mustapha, Anas Al Bassit, and Sabri Skhiri. Towards Privacy

Policy Conceptual Modeling. In Gillian Dobbie, Ulrich Frank, Gerti Kappel, Stephen W. Liddle, and Heinrich C. Mayr, editors, *Conceptual Modeling*, Lecture Notes in Computer Science, pages 429–438. Springer International Publishing, 2020. ISBN 978-3-030-62522-1. doi: 10.1007/978-3-030-62522-1_32.

P.B. Kruchten. The 4+1 View Model of architecture. *IEEE Software*, 12(6):42–50, 1995. ISSN 1937-4194. doi: 10.1109/52.469759.

Ponnurangam Kumaraguru and Lorrie Faith Cranor. Privacy Indexes: A Survey of Westin's Studies. Technical report, 2005.

Jonathan Lawson, Moran N. Cabili, Giselle Kerry, Tiffany Boughtwood, Adrian Thorogood, Pinar Alper, Sarion R. Bowers, Rebecca R. Boyles, Anthony J. Brookes, Matthew Brush, Tony Burdett, Hayley Clissold, Stacey Donnelly, Stephanie O. M. Dyke, Mallory A. Freeberg, Melissa A. Haendel, Chihiro Hata, Petr Holub, Francis Jeanson, Aina Jene, Minae Kawashima, Shuichi Kawashima, Melissa Konopko, Irene Kyomugisha, Haoyuan Li, Mikael Linden, Laura Lyman Rodriguez, Mizuki Morita, Nicola Mulder, Jean Muller, Satoshi Nagaie, Jamal Nasir, Soichi Ogishima, Vivian Ota Wang, Laura D. Paglione, Ravi N. Pandya, Helen Parkinson, Anthony A. Philippakis, Fabian Prasser, Jordi Rambla, Kathy Reinold, Gregory A. Rushton, Andrea Saltzman, Gary Saunders, Heidi J. Sofia, John D. Spalding, Morris A. Swertz, Ilia Tulchinsky, Esther J. van Enckevort, Susheel Varma, Craig Voisin, Natsuko Yamamoto, Chisato Yamasaki, Lyndon Zass, Jaime M. Guidry Auvil, Tommi H. Nyrönen, and Mélanie Courtot. The Data Use Ontology to streamline responsible access to human biomedical datasets. *Cell Genomics*, 1(2), 2021. ISSN 2666-979X. doi: 10.1016/j.xgen.2021.100028.

Daniel Le Métayer and Shara Monteleone. Automated consent through privacy agents: Legal requirements and technical architecture. *Computer Law & Security Review*, 25(2):136–144, 2009. ISSN 0267-3649. URL https://www.sciencedirect.com/science/article/pii/S0267364909000387.

Timothy Lebo, Satya Sahoo, and Deborah McGuinness. PROV-O: The PROV Ontology – W3C Recommendation 30 April 2013, 2013. URL https://www.w3.org/TR/prov-o/.

Jens Leicht and Maritta Heisel. A Survey on Privacy Policy Languages: Expressiveness Concerning Data Protection Regulations. In *2019 12th CMI Conference on Cybersecurity and Privacy (CMI)*, pages 1–6. IEEE, 2019. ISBN 978-1-72812-856-6. doi: 10.1109/CMI48017.2019.8962144.

Ninghui Li, Ting Yu, and Annie Antón. A semantics-base approach to privacy languages. *Computer Systems: Science & Engineering - CSSE*, 21, 2006.

Thomas Linden, Rishabh Khandelwal, Hamza Harkous, and Kassem Fawaz. The Privacy Policy Landscape After the GDPR. In *Proceedings on Privacy Enhancing Technologies*, volume 1, pages 47–64, 2020. doi: 10.2478/popets-2020-0004.

Georgios Lioudakis and Davide Cascone. Compliance Ontology – Deliverable D3.1. Project deliverable, 2019. URL https://www.bpr4gdpr.eu/wp-content/uploads/2019/06/D3.1-Compliance-Ontology-1.0.pdf.

Juniper Lovato, Philip Mueller, Parisa Suchdev, and Peter Dodds. More Data Types More Problems: A Temporal Analysis of Complexity, Stability, and Sensitivity in Privacy Policies. In *2023 ACM*

Conference on Fairness, Accountability, and Transparency, pages 1088–1100. Association for Computing Machinery, 2023. ISBN 9798400701924. doi: 10.1145/3593013.3594065.

Bernadette Farias Lóscio, Caroline Burle, and Newton Calegari. Data on the Web Best Practices – W3C Recommendation 31 January 2017, 2017. URL https://www.w3.org/TR/dwbp/.

Essam Mansour, Andrei Vlad Sambra, Sandro Hawke, Maged Zereba, Sarven Capadisli, Abdurrahman Ghanem, Ashraf Aboulnaga, and Tim Berners-Lee. A Demonstration of the Solid Platform for Social Web Applications. In *Proceedings of the 25th International Conference Companion on World Wide Web*, WWW '16 Companion, pages 223–226. International World Wide Web Conferences Steering Committee, 2016. ISBN 978-1-4503-4144-8. doi: 10.1145/2872518.2890529.

Marcello M. Mariani, Maria Ek Styven, and Fréderic Teulon. Explaining the intention to use digital personal data stores: An empirical study. *Technological Forecasting and Social Change*, 166, 2021. ISSN 0040-1625. doi: 10.1016/j.techfore.2021.120657.

Paula Andrea Martinez, Christopher Erdmann, Natasha Simons, Reid Otsuji, Stephanie Labou, Ryan Johnson, Guilherme Castelao, Bia Villas Boas, Anna-Lena Lamprecht, Carlos Martinez Ortiz, Leyla Garcia, Mateusz Kuzak, Liz Stokes, Tom Honeyman, Sharyn Wise, Josh Quan, Scott Peterson, Amy Neeser, Lena Karvovskaya, Otto Lange, Iza Witkowska, Jacques Flores, Fiona Bradley, Kristina Hettne, Peter Verhaar, Ben Companjen, Laurents Sesink, Fieke Schoots, Erik Schultes, Rajaram Kaliyaperumal, Erzsébet Tóth-Czifra, Ricardo de Miranda Azevedo, Sanne Muurling, John Brown, Janice Chan, Niamh Quigley, Lisa Federer, Douglas Joubert, Allissa Dillman, Kenneth Wilkins, Ishwar Chandramouliswaran, Vivek Navale, Susan Wright, Silvia Di Giorgio, Mandela Fasemore, Konrad Förstner, Till Sauerwein, Eva Seidlmayer, Ilja Zeitlin, Susannah Bacon, Katie Hannan, Richard Ferrers, Keith Russell, Deidre Whitmore, Tim Dennis, Daniel Bangert, Albert Meroño Peñuela, Enrico Daga, Gerry Ryder, Aswin Narayanan, Iryna Kuchma, Jose Manzano Patron, Andrew Mehnert, Matthias Liffers, Ronald Siebes, Gerard Coen, Kathleen Gregory, Andrea Scharnhorst, Maria Cruz, Francoise Genova, Matthew Kenworthy, Natalie Meyers, Evert Rol, Juande Santander-Vela, Joanne Yeomans, Elli Papadopoulou, Emma Lazzeri, Leonidas Mouchliadis, Katerina Lenaki, Spyros Zoupanos, Danail Hristozov, Stella Stoycheva, Ellen Leenarts, Marjan Grootveld, Frans Huigen, and Eliane Fankhauser. Top 10 FAIR Data & Software Things. Technical report, 2019. URL https://zenodo.org/records/3409968.

Karsten Martiny and Grit Denker. Partial Decision Overrides in a Declarative Policy Framework. In *2020 IEEE 14th International Conference on Semantic Computing (ICSC)*, pages 271–278. IEEE, 2020. ISBN 978-1-72816-332-1. doi: 10.1109/ICSC.2020.00056.

Karsten Martiny, Daniel Elenius, and Grit Denker. Protecting Privacy with a Declarative Policy Framework. In *2018 IEEE 12th International Conference on Semantic Computing (ICSC)*, pages 227–234. IEEE, 2018. ISBN 978-1-5386-4408-9. doi: 10.1109/ICSC.2018.00039.

Aleecia M. McDonald and Lorrie Faith Cranor. The Cost of Reading Privacy Policies. *I/S: A Journal of Law and Policy for the Information Society*, 4(3):543–568, 2008.

Peter Mechant, Ralf De Wolf, Mathias Van Compernolle, Glen Joris, Tom Evens, and Lieven De Marez. Saving the web by decentralizing data networks? A socio-technical reflection on the promise of decentralization and personal data stores. In *2021 14th CMI International Conference -*

Critical ICT Infrastructures and Platforms (CMI), pages 1–6, 2021. doi: 10.1109/CMI53512.2021.9663788.

Jayashree Mohan, Melissa Wasserman, and Vijay Chidambaram. Analyzing GDPR Compliance Through the Lens of Privacy Policy. In Vijay Gadepally, Timothy Mattson, Michael Stonebraker, Fusheng Wang, Gang Luo, Yanhui Laing, and Alevtina Dubovitskaya, editors, *Heterogeneous Data Management, Polystores, and Analytics for Healthcare*, Lecture Notes in Computer Science, pages 82–95. Springer International Publishing, 2019. ISBN 978-3-030-33752-0. doi: 10.1007/978-3-030-33752-0_6.

Jackson Morgan, Aaron Coburn, and Matthieu Bosquet. Solid-OIDC Primer Version 0.1.0. *W3C Community Group Draft Report*, 2022. URL https://solidproject.org/TR/oidc-primer.

Victoria Nembaware, Katherine Johnston, Alpha A. Diallo, Maritha J. Kotze, Alice Matimba, Keymanthri Moodley, Godfrey B. Tangwa, Rispah Torrorey-Sawe, and Nicki Tiffin. A framework for tiered informed consent for health genomic research in Africa. *Nature Genetics*, 51(11): 1566–1571, 2019. ISSN 1546-1718. doi: 10.1038/s41588-019-0520-x.

Helen Nissenbaum. Privacy as Contextual Integrity. *Washington Law Review*, 79(1):119, 2004. URL https://digitalcommons.law.uw.edu/wlr/vol79/iss1/10.

Whitney Nixdorf. Planting in a Walled Garden: Data Portability Policies to Inform Consumers How Much (If Any) of the Harvest Is Their Share. *Transnational Law and Contemporary Problems*, 29: 135–164, 2019.

Jonathan A. Obar and Anne Oeldorf-Hirsch. The biggest lie on the Internet: ignoring the privacy policies and terms of service policies of social networking services. *Information, Communication & Society*, 23(1):128–147, 2020. ISSN 1369-118X. doi: 10.1080/1369118X.2018.1486870.

Office of Publications on Eur-Lex. EU Vocabularies – European Legislation Identifier (ELI), 2017. URL https://op.europa.eu/en/web/eu-vocabularies/eli.

Monica Palmirani and Guido Governatori. Modelling Legal Knowledge for GDPR Compliance Checking. In *Legal Knowledge and Information Systems*, volume 313, pages 101–110, 2018. doi: 10.3233/978-1-61499-935-5-101.

Monica Palmirani and Fabio Vitali. Akoma-Ntoso for Legal Documents. In Giovanni Sartor, Monica Palmirani, Enrico Francesconi, and Maria Angela Biasiotti, editors, *Legislative XML for the Semantic Web: Principles, Models, Standards for Document Management*, Law, Governance and Technology Series, pages 75–100. Springer Netherlands, 2011. ISBN 978-94-007-1887-6. doi: 10.1007/978-94-007-1887-6_6.

Monica Palmirani, Michele Martoni, Arianna Rossi, Cesare Bartolini, and Livio Robaldo. PrOnto: Privacy Ontology for Legal Reasoning. In Andrea Kő and Enrico Francesconi, editors, *Electronic Government and the Information Systems Perspective (EGOVIS 2018)*, volume 11032 of *Lecture Notes in Computer Science*, pages 139–152. Springer International Publishing, 2018. ISBN 978-3-319-98348-6 978-3-319-98349-3. doi: 10.1007/978-3-319-98349-3_11.

Monica Palmirani, Guido Governatori, Tara Athan, Harold Boley, Adrian Paschke, and

Adam Wyner. LegalRuleML Core Specification Version 1.0 – OASIS Standard, 2021. URL https://docs.oasis-open.org/legalruleml/legalruleml-core-spec/v1.0/os/legalruleml-core-spec-v1.0-os.html.

Harshvardhan J. Pandit. Representing Activities associated with Processing of Personal Data and Consent using Semantic Web for GDPR Compliance. *Doctoral Thesis. Trinity College Dublin*, 2020.

Harshvardhan J. Pandit. Making Sense of Solid for Data Governance and GDPR. *Information*, 14 (2), 2023. ISSN 2078-2489. doi: 10.3390/info14020114.

Harshvardhan J. Pandit and Beatriz Esteves. Enhancing Data Use Ontology (DUO) for Health-Data Sharing by Extending it with ODRL and DPV. *Semantic Web Journal*, 2024. doi: 10.3233/SW-243583.

Harshvardhan J. Pandit and Dave Lewis. Modelling Provenance for GDPR Compliance using Linked Open Data Vocabularies. In *Society, Privacy and the Semantic Web – Policy and Technology (PrivOn 2017), co-located with ISWC 2017*, volume 1951, 2017. URL http://ceur-ws.org/Vol-1951/PrivOn2017_paper_6.pdf.

Harshvardhan J. Pandit, Kaniz Fatema, Declan O'Sullivan, and Dave Lewis. GDPRtEXT – GDPR as a Linked Data Resource. In Aldo Gangemi, Roberto Navigli, Maria-Esther Vidal, Pascal Hitzler, Raphaël Troncy, Laura Hollink, Anna Tordai, and Mehwish Alam, editors, *The Semantic Web*, volume 10843 of *Lecture Notes in Computer Science*, pages 481–495. Springer International Publishing, 2018. ISBN 978-3-319-93416-7 978-3-319-93417-4. doi: 10.1007/978-3-319-93417-4_31.

Harshvardhan J. Pandit, Christophe Debruyne, Declan O'Sullivan, and Dave Lewis. GConsent – A Consent Ontology Based on the GDPR. In Pascal Hitzler, Miriam Fernández, Krzysztof Janowicz, Amrapali Zaveri, Alasdair J.G. Gray, Vanessa Lopez, Armin Haller, and Karl Hammar, editors, *The Semantic Web*, volume 11503 of *Lecture Notes in Computer Science*, pages 270–282. Springer International Publishing, 2019a. ISBN 978-3-030-21347-3 978-3-030-21348-0. doi: 10.1007/978-3-030-21348-0_18.

Harshvardhan J. Pandit, Axel Polleres, Bert Bos, Rob Brennan, Bud Bruegger, Fajar J. Ekaputra, Javier D. Fernández, Roghaiyeh Gachpaz Hamed, Elmar Kiesling, Mark Lizar, Eva Schlehahn, Simon Steyskal, and Rigo Wenning. Creating a Vocabulary for Data Privacy: The First-Year Report of Data Privacy Vocabularies and Controls Community Group (DPVCG). In Hervé Panetto, Christophe Debruyne, Martin Hepp, Dave Lewis, Claudio Agostino Ardagna, and Robert Meersman, editors, *On the Move to Meaningful Internet Systems: OTM 2019 Conferences*, volume 11877 of *Lecture Notes in Computer Science*, pages 714–730. Springer International Publishing, 2019b. ISBN 978-3-030-33245-7 978-3-030-33246-4. doi: 10.1007/978-3-030-33246-4_44.

Harshvardhan J. Pandit, Georg P. Krog, and Beatriz Esteves. Primer Data Privacy Vocabulary (DPV) – Final Community Group Report 05 December 2022, 2022. URL https://w3id.org/dpv/primer.

Eugenia I. Papagiannakopoulou, Maria N. Koukovini, Georgios Lioudakis, Nikolaos Dellas, Joaquin Garcia-Alfaro, Dimitra I. Kaklamani, Iakovos S. Venieris, Nora Cuppens-Boulahia, and Frédéric Cuppens. Leveraging Ontologies upon a Holistic Privacy-Aware Access Control Model. In J. Dan-

ger, M. Debbabi, JY. Marion, J. Garcia-Alfaro, and N. Zincir Heywood, editors, *Foundations and Practice of Security. FPS 2013*, volume 8352 of *Lecture Notes in Computer Science*, pages 209–226. Springer, Cham, 2014. URL https://www.researchgate.net/publication/260706010_Leveraging_Ontologies_upon_a_Holistic_Privacy-Aware_Access_Control_Model.

Bill Parducci, Hal Lockhart, and Erik Rissanen. eXtensible Access Control Markup Language (XACML) Version 3.0 – OASIS Standard, 2013. URL http://docs.oasis-open.org/xacml/3.0/xacml-3.0-core-spec-os-en.html.

Frank Pasquale. *The Black Box Society: The Secret Algorithms That Control Money and Information*. Harvard University Press, 2015. ISBN 978-0-674-36827-9. doi: 10.4159/harvard.9780674736061.

PCI Security Standards Council. Payment Card Industry (PCI) Data Security Standard – Version 3.2.1, 2018. URL https://www.pcisecuritystandards.org/document_library.

Tassilo Pellegrini, Victor Mireles, Simon Steyskal, Oleksandra Panasiuk, Anna Fensel, and Sabrina Kirrane. Automated Rights Clearance Using Semantic Web Technologies: The DALICC Framework. In Thomas Hoppe, Bernhard Humm, and Anatol Reibold, editors, *Semantic Applications: Methodology, Technology, Corporate Use*, pages 203–218. Springer, Berlin, Heidelberg, 2018a. ISBN 978-3-662-55433-3. doi: 10.1007/978-3-662-55433-3_14.

Tassilo Pellegrini, Andrea Schönhofer, Sabrina Kirrane, Anna Fensel, Oleksandra Panasiuk, Victor Mireles-Chavez, Thomas Thurner, Markus Dörfler, and Axel Polleres. A Genealogy and Classification of Rights Expression Languages - Preliminary Results. In *Proceedings of the 21st International Legal Informatics Symposium*, pages 243–250, 2018b.

Silvio Peroni. The Semantic Publishing and Referencing Ontologies. In *Semantic Web Technologies and Legal Scholarly Publishing*, volume 15 of *Law, Governance and Technology Series*, pages 121–193. Springer, Cham, 2014. ISBN 978-3-319-04776-8.

Luca Piras, Mohammed Ghazi Al-Obeidallah, Andrea Praitano, Aggeliki Tsohou, Haralambos Mouratidis, Beatriz Gallego-Nicasio Crespo, Jean Baptiste Bernard, Marco Fiorani, Emmanouil Magkos, Andrès Castillo Sanz, Michalis Pavlidis, Roberto D'Addario, and Giuseppe Giovanni Zorzino. DEFeND Architecture: A Privacy by Design Platform for GDPR Compliance. In Stefanos Gritzalis, Edgar R. Weippl, Sokratis K. Katsikas, Gabriele Anderst-Kotsis, A Min Tjoa, and Ismail Khalil, editors, *Trust, Privacy and Security in Digital Business*, Lecture Notes in Computer Science, pages 78–93. Springer International Publishing, 2019. ISBN 978-3-030-27813-7. doi: 10.1007/978-3-030-27813-7_6.

María Poveda-Villalón, Asunción Gómez-Pérez, and Mari Carmen Suárez-Figueroa. OOPS! (OntOlogy Pitfall Scanner!): An On-line Tool for Ontology Evaluation. *International Journal on Semantic Web and Information Systems*, 10(2):7–34, 2014. ISSN 1552-6283. doi: 10.4018/ijswis.2014040102.

María Poveda-Villalón, Alba Fernández-Izquierdo, Mariano Fernández-López, and Raúl García-Castro. LOT: An industrial oriented ontology engineering framework. *Engineering Applications of Artificial Intelligence*, 111:104755, 2022. ISSN 0952-1976. doi: 10.1016/j.engappai.2022.104755.

Privacy Level Agreement Working Group. Code of Conduct for GDPR Compliance,

2017. URL https://downloads.cloudsecurityalliance.org/assets/research/gdpr/CSA_Code_of_Conduct_for_GDPR_Compliance.pdf.

Eric Prud'hommeaux and Gavin Carothers. RDF 1.1 Turtle: Terse RDF Triple Language. *W3C Recommendation*, 2014. URL https://www.w3.org/TR/turtle/.

Eric Prud'hommeaux, Iovka Boneva, Jose Emilio Labra Gayo, and Gregg Kellogg. Shape Expressions Language 2.1 – final Community Group Report 8 October 2019, 2019. URL http://shex.io/shex-semantics/.

Publications Office of the European Union. Named Authority List: Access rights, 2023. URL http://publications.europa.eu/resource/dataset/access-right.

Heidi L. Rehm, Angela J. H. Page, Lindsay Smith, Jeremy B. Adams, Gil Alterovitz, Lawrence J. Babb, Maxmillian P. Barkley, Michael Baudis, Michael J. S. Beauvais, Tim Beck, Jacques S. Beckmann, Sergi Beltran, David Bernick, Alexander Bernier, James K. Bonfield, Tiffany F. Boughtwood, Guillaume Bourque, Sarion R. Bowers, Anthony J. Brookes, Michael Brudno, Matthew H. Brush, David Bujold, Tony Burdett, Orion J. Buske, Moran N. Cabili, Daniel L. Cameron, Robert J. Carroll, Esmeralda Casas-Silva, Debyani Chakravarty, Bimal P. Chaudhari, Shu Hui Chen, J. Michael Cherry, Justina Chung, Melissa Cline, Hayley L. Clissold, Robert M. Cook-Deegan, Mélanie Courtot, Fiona Cunningham, Miro Cupak, Robert M. Davies, Danielle Denisko, Megan J. Doerr, Lena I. Dolman, Edward S. Dove, L. Jonathan Dursi, Stephanie O. M. Dyke, James A. Eddy, Karen Eilbeck, Kyle P. Ellrott, Susan Fairley, Khalid A. Fakhro, Helen V. Firth, Michael S. Fitzsimons, Marc Fiume, Paul Flicek, Ian M. Fore, Mallory A. Freeberg, Robert R. Freimuth, Lauren A. Fromont, Jonathan Fuerth, Clara L. Gaff, Weiniu Gan, Elena M. Ghanaim, David Glazer, Robert C. Green, Malachi Griffith, Obi L. Griffith, Robert L. Grossman, Tudor Groza, Jaime M. Guidry Auvil, Roderic Guigó, Dipayan Gupta, Melissa A. Haendel, Ada Hamosh, David P. Hansen, Reece K. Hart, Dean Mitchell Hartley, David Haussler, Rachele M. Hendricks-Sturrup, Calvin W. L. Ho, Ashley E. Hobb, Michael M. Hoffman, Oliver M. Hofmann, Petr Holub, Jacob Shujui Hsu, Jean-Pierre Hubaux, Sarah E. Hunt, Ammar Husami, Julius O. Jacobsen, Saumya S. Jamuar, Elizabeth L. Janes, Francis Jeanson, Aina Jené, Amber L. Johns, Yann Joly, Steven J. M. Jones, Alexander Kanitz, Kazuto Kato, Thomas M. Keane, Kristina Kekesi-Lafrance, Jerome Kelleher, Giselle Kerry, Seik-Soon Khor, Bartha M. Knoppers, Melissa A. Konopko, Kenjiro Kosaki, Martin Kuba, Jonathan Lawson, Rasko Leinonen, Stephanie Li, Michael F. Lin, Mikael Linden, Xianglin Liu, Isuru Udara Liyanage, Javier Lopez, Anneke M. Lucassen, Michael Lukowski, Alice L. Mann, John Marshall, Michele Mattioni, Alejandro Metke-Jimenez, Anna Middleton, Richard J. Milne, Fruzsina Molnár-Gábor, Nicola Mulder, Monica C. Munoz-Torres, Rishi Nag, Hidewaki Nakagawa, Jamal Nasir, Arcadi Navarro, Tristan H. Nelson, Ania Niewielska, Amy Nisselle, Jeffrey Niu, Tommi H. Nyrönen, Brian D. O'Connor, Sabine Oesterle, Soichi Ogishima, Vivian Ota Wang, Laura A. D. Paglione, Emilio Palumbo, Helen E. Parkinson, Anthony A. Philippakis, Angel D. Pizarro, Andreas Prlic, Jordi Rambla, Augusto Rendon, Renee A. Rider, Peter N. Robinson, Kurt W. Rodarmer, Laura Lyman Rodriguez, Alan F. Rubin, Manuel Rueda, Gregory A. Rushton, Rosalyn S. Ryan, Gary I. Saunders, Helen Schuilenburg, Torsten Schwede, Serena Scollen, Alexander Senf, Nathan C. Sheffield, Neerjah Skantharajah, Albert V. Smith, Heidi J. Sofia, Dylan Spalding, Amanda B. Spurdle, Zornitza Stark, Lincoln D. Stein, Makoto Suematsu, Patrick Tan, Jonathan A. Tedds, Alastair A. Thomson, Adrian Thorogood, Timothy L. Tickle, Katsushi Tokunaga, Juha Törnroos, David Torrents, Sean Upchurch, Alfonso Valencia, Roman Valls Guimera, Jessica

Vamathevan, Susheel Varma, Danya F. Vears, Coby Viner, Craig Voisin, Alex H. Wagner, Susan E. Wallace, Brian P. Walsh, Marc S. Williams, Eva C. Winkler, Barbara J. Wold, Grant M. Wood, J. Patrick Woolley, Chisato Yamasaki, Andrew D. Yates, Christina K. Yung, Lyndon J. Zass, Ksenia Zaytseva, Junjun Zhang, Peter Goodhand, Kathryn North, and Ewan Birney. GA4GH: International policies and standards for data sharing across genomic research and healthcare. *Cell Genomics*, 1(2), 2021. ISSN 2666-979X. doi: 10.1016/j.xgen.2021.100029.

Viktoria H. S. E. Robertson. Excessive data collection: Privacy considerations and abuse of dominance in the era of big data. *Common Market Law Review*, 57(1), 2020. ISSN 0165-0750. doi: 10.54648/cola2020006.

Manuel Rudolph, Denis Feth, and Svenja Polst. Why Users Ignore Privacy Policies – A Survey and Intention Model for Explaining User Privacy Behavior. In Masaaki Kurosu, editor, *Human-Computer Interaction. Theories, Methods, and Human Issues*, pages 587–598, Cham, 2018. Springer International Publishing. ISBN 978-3-319-91238-7. doi: 10.1007/978-3-319-91238-7_45.

John M. M. Rumbold and Barbara K. Pierscionek. What Are Data? A Categorization of the Data Sensitivity Spectrum. *Big Data Research*, 12:49–59, 2018. ISSN 2214-5796. doi: 10.1016/j.bdr.2017.11.001.

Owen Sacco and Alexandre Passant. A Privacy Preference Ontology (PPO) for Linked Data. 2011a. URL http://ceur-ws.org/Vol-813/ldow2011-paper01.pdf.

Owen Sacco and Alexandre Passant. A Privacy Preference Manager for the Social Semantic Web. In *Proceedings of the 2nd Workshop on Semantic Personalized Information Management: Retrieval and Recommendation, SPIM2011*, pages 42–53, 2011b. ISBN 16130073.

Nat Sakimura, John Bradley, Michael B. Jones, Breno de Medeiros, and Chuck Mortimore. OpenID Connect Core 1.0. Technical report, 2014. URL https://openid.net/specs/openid-connect-core-1_0.html.

Andrei Vlad Sambra, Essam Mansour, Sandro Hawke, Maged Zereba, Nicola Greco, Abdurrahman Ghanem, Dmitri Zagidulin, Ashraf Aboulnaga, and Tim Berners-Lee. Solid: A platform for decentralized social applications based on linked data. Technical report, 2016.

Cristiana Santos and Harshvardhan J. Pandit. How could the upcoming ePrivacy Regulation recognise enforceable privacy signals in the EU?, 2023. URL https://doi.org/10.31219/osf.io/xvyf3.

Gary Saunders, Michael Baudis, Regina Becker, Sergi Beltran, Christophe Béroud, Ewan Birney, Cath Brooksbank, Søren Brunak, Marc Van den Bulcke, Rachel Drysdale, Salvador Capella-Gutierrez, Paul Flicek, Francesco Florindi, Peter Goodhand, Ivo Gut, Jaap Heringa, Petr Holub, Jef Hooyberghs, Nick Juty, Thomas M. Keane, Jan O. Korbel, Ilkka Lappalainen, Brane Leskosek, Gert Matthijs, Michaela Th Mayrhofer, Andres Metspalu, Arcadi Navarro, Steven Newhouse, Tommi Nyrönen, Angela Page, Bengt Persson, Aarno Palotie, Helen Parkinson, Jordi Rambla, David Salgado, Erik Steinfelder, Morris A. Swertz, Alfonso Valencia, Susheel Varma, Niklas Blomberg, and Serena Scollen. Leveraging European infrastructures to access 1 million human genomes by 2022. *Nature Reviews Genetics*, 20(11):693–701, 2019. ISSN 1471-0064. doi: 10.1038/s41576-019-0156-9.

Michael C. Schatz, Anthony A. Philippakis, Enis Afgan, Eric Banks, Vincent J. Carey, Robert J. Carroll, Alessandro Culotti, Kyle Ellrott, Jeremy Goecks, Robert L. Grossman, Ira M. Hall, Kasper D. Hansen, Jonathan Lawson, Jeffrey T. Leek, Anne O'Donnell Luria, Stephen Mosher, Martin Morgan, Anton Nekrutenko, Brian D. O'Connor, Kevin Osborn, Benedict Paten, Candace Patterson, Frederick J. Tan, Casey Overby Taylor, Jennifer Vessio, Levi Waldron, Ting Wang, and Kristin Wuichet. Inverting the model of genomics data sharing with the NHGRI Genomic Data Science Analysis, Visualization, and Informatics Lab-space. *Cell Genomics*, 2(1), 2022. ISSN 2666-979X. doi: 10.1016/j.xgen.2021.100085.

Sara Shakeri, Valentina Maccatrozzo, Lourens Veen, Rena Bakhshi, Leon Gommans, Cees de Laat, and Paola Grosso. Modeling and Matching Digital Data Marketplace Policies. In *2019 15th International Conference on eScience (eScience)*, pages 570–577, 2019. doi: 10.1109/eScience.2019.00078.

Mark Sheehan. Can Broad Consent be Informed Consent? *Public Health Ethics*, 4(3):226–235, 2011. ISSN 1754-9973. doi: 10.1093/phe/phr020.

Wout Slabbinck, Ruben Dedecker, Julián Andrés Rojas, and Ruben Verborgh. A Rule-Based Software Agent on Top of Personal Data Stores. In *ISWC 2023 Posters and Demos*, volume 3632 of *22nd International Semantic Web Conference*. CEUR Workshop Proceedings, 2023. URL https://ceur-ws.org/Vol-3632/ISWC2023_paper_406.pdf.

Benedict Whittam Smith, Víctor Rodríguez-Doncel, and Beatriz Esteves. ODRL Implementation Best Practices – Draft Community Group Report 28 July 2023, 2023. URL https://w3c.github.io/odrl/bp/.

James M. Snell and Evan Prodromou. Activity Streams 2.0 – W3C Recommendation 23 May 2017, 2017. URL https://www.w3.org/TR/activitystreams-core/.

Hannah Snyder. Literature review as a research methodology: An overview and guidelines. *Journal of Business Research*, 104:333–339, 2019. ISSN 0148-2963. doi: 10.1016/j.jbusres.2019.07.039.

Solid Editorial Team. Use Cases and Requirements for Authorization in Solid. *W3C Community Group Draft Report*, 2023. URL https://solid.github.io/authorization-panel/authorization-ucr/.

Daniel J. Solove. Privacy Self-Management and the Consent Dilemma. *Harvard Law Review*, 126, 2012. URL https://papers.ssrn.com/abstract=2171018.

Daniel J. Solove. Murky Consent: An Approach to the Fictions of Consent in Privacy Law. *104 Boston University Law Review (Forthcoming)*, 2023. doi: 10.2139/ssrn.4333743.

Steve Speicher, John Arwe, and Ashok Malhotra. Linked Data Platform 1.0. *W3C Recommendation*, 2015. URL https://www.w3.org/TR/ldp/.

Manu Sporny, Dave Longley, Markus Sabadello, Drummond Reed, Orie Steele, and Christopher Allen. Decentralized Identifiers (DIDs) v1.0: Core architecture, data model, and representations. *W3C Recommendation*, 2022. URL https://www.w3.org/TR/did-core/.

Manu Sporny, Dave Longley, and David Chadwick. Verifiable Credentials Data Model v2.0. *W3C Working Draft*, 2023. URL https://www.w3.org/TR/vc-data-model-2.0/.

Michael Steidl. ODRL Profile Best Practices – Draft Community Group Report 28 July 2023, 2023. URL https://w3c.github.io/odrl/profile-bp/.

Henry Story, Stéphane Corlosquet, and Andrei Sambra. WebID-TLS: WebID Authentication over TLS. *W3C Editor's Draft*, 2014. URL https://www.w3.org/2005/Incubator/webid/spec/tls/.

Chang Sun, Marc Gallofré Ocaña, Johan van Soest, and Michel Dumontier. ciTIzen-centric DAta pLatform (TIDAL): Sharing distributed personal data in a privacy-preserving manner for health research. *Semantic Web*, 14(5):977–996, 2023. ISSN 1570-0844. doi: 10.3233/SW-223220. Publisher: IOS Press.

Mari Carmen Suárez-Figueroa, Asunción Gómez-Pérez, and Mariano Fernández-López. The NeOn Methodology for Ontology Engineering. In Mari Carmen Suárez-Figueroa, Asunción Gómez-Pérez, Enrico Motta, and Aldo Gangemi, editors, *Ontology Engineering in a Networked World*, pages 9–34. Springer Berlin Heidelberg, 2012. ISBN 978-3-642-24794-1. doi: 10.1007/978-3-642-24794-1_2.

Ruben Taelman, Joachim Van Herwegen, Miel Vander Sande, and Ruben Verborgh. Comunica: A modular SPARQL query engine for the web. In Denny Vrandečić, Kalina Bontcheva, Mari Carmen Suárez-Figueroa, Valentina Presutti, Irene Celino, Marta Sabou, Lucie-Aimée Kaffee, and Elena Simperl, editors, *The Semantic Web – ISWC 2018*, Lecture Notes in Computer Science, pages 239–255. Springer International Publishing, 2018. ISBN 978-3-030-00668-6. doi: 10.1007/978-3-030-00668-6_15.

Humphrey Taylor. Most People Are "Privacy Pragmatists" Who, While Concerned about Privacy, Will Sometimes Trade It Off for Other Benefits. Technical report, 2003. Conducted among 1,010 respondents.

Arnout Terpstra, Alexander P. Schouten, Alwin de Rooij, and Ronald E. Leenes. Improving privacy choice through design: How designing for reflection could support privacy self-management. *First Monday*, 24(7), 2019. ISSN 1396-0466. doi: 10.5210/fm.v24i7.9358.

Gergely Tóth. Preserving control over user data in the hospitality industry with Solid. *Master's thesis*, 2022. URL https://dspace.cuni.cz/handle/20.500.11956/176270. Charles University.

UK Government. Data: a new direction – government response to consultation, 2022. URL https://www.gov.uk/government/consultations/data-a-new-direction/outcome/data-a-new-direction-government-response-to-consultation.

United Nations General Assembly. Universal Declaration of Human Rights, 1948. URL https://www.un.org/en/about-us/universal-declaration-of-human-rights.

Sander Van Damme, Peter Mechant, Eveline Vlassenroot, Mathias Van Compernolle, Raf Buyle, and Dorien Bauwens. Towards a Research Agenda for Personal Data Spaces: Synthesis of a Community Driven Process. In Marijn Janssen, Csaba Csáki, Ida Lindgren, Euripidis Loukis, Ulf Melin, Gabriela Viale Pereira, Manuel Pedro Rodríguez Bolívar, and Efthimios Tambouris,

editors, *Electronic Government*, Lecture Notes in Computer Science, pages 563–577. Springer International Publishing, 2022. ISBN 978-3-031-15086-9. doi: 10.1007/978-3-031-15086-9_36.

Maxim Van de Wynckel and Beat Signer. A Solid-based Architecture for Decentralised Interoperable Location Data. In *12th International Conference on Indoor Positioning and Indoor Navigation (IPIN 2022)*, volume 3248 of *CEUR Workshop Proceedings*, 2022. URL https://ceur-ws.org/Vol-3248/paper11.pdf.

Pierre-Yves Vandenbussche, Ghislain A. Atemezing, María Poveda-Villalón, and Bernard Vatant. Linked Open Vocabularies (LOV): A gateway to reusable semantic vocabularies on the Web. *Semantic Web*, 8(3):437–452, 2017. ISSN 1570-0844. doi: 10.3233/SW-160213.

Ruben Verborgh. Paradigm shifts for the decentralized Web, 2017. URL https://ruben.verborgh.org/blog/2017/12/20/paradigm-shifts-for-the-decentralized-web/.

Ruben Verborgh. Re-decentralizing the Web, for good this time. In Oshani Seneviratne and James Hendler, editors, *Linking the World's Information: A Collection of Essays on the Work of Sir Tim Berners-Lee*. ACM, 2022. URL https://ruben.verborgh.org/articles/redecentralizing-the-web/.

Ruben Verborgh. No more raw data, 2023. URL https://ruben.verborgh.org/blog/2023/11/10/no-more-raw-data/.

Sofie Verbrugge, Frederic Vannieuwenborg, Marlies Van der Wee, Didier Colle, Ruben Taelman, and Ruben Verborgh. Towards a personal data vault society: an interplay between technological and business perspectives. In *2021 60th FITCE Communication Days Congress for ICT Professionals: Industrial Data — Cloud, Low Latency and Privacy (FITCE)*, pages 1–6, 2021. doi: 10.1109/FITCE53297.2021.9588540.

Melanie Verstraete, Sofie Verbrugge, and Didier Colle. Solid: Enabler of decentralized, digital platforms ecosystems. In *31st European Conference of the International Telecommunications Society (ITS): "Reining in Digital Platforms? Challenging monopolies, promoting competition and developing regulatory regimes"*, 2022. URL http://hdl.handle.net/10419/265673.

Salomé Viljoen. A Relational Theory of Data Governance. *Yale Law Journal*, 131(2):573–654, 2021. URL https://www.yalelawjournal.org/feature/a-relational-theory-of-data-governance.

Angela Vivarelli. The Crisis of the Right to Informational Self-Determination. *Italian Law Journal*, 6(1):301–319, 2020. URL https://theitalianlawjournal.it/data/uploads/6-italj-1-2020/301-vivarelli.pdf.

Yannick Alexander Vogel. Stretching the Limit, the Functioning of the GDPR's Notion of Consent in the Context of Data Intermediary Services. *European Data Protection Law Review*, 8(2):238–249, 2022. URL https://heinonline.org/HOL/P?h=hein.journals/edpl8&i=246.

Ari Ezra Waldman. *Industry Unbound: The Inside Story of Privacy, Data, and Corporate Power*. Cambridge University Press, 1st edition, 2021. ISBN 978-1-108-59138-6. doi: 10.1017/9781108591386.

Lisa Wang. Enhancing Public Service Delivery by improving Governmental Processes through the decentralized Solid Ecosystem. *Master's thesis*, 2020. URL https://rog.pleio.nl/file/download/a692d988-963d-4dbc-908f-1d7aae37f845/1605087634thesis_solid.pdf. Vrije Universiteit Amsterdam.

Jane Webster and Richard T. Watson. Analyzing the Past to Prepare for the Future: Writing a Literature Review. *MIS Quarterly*, 26(2):xiii–xxiii, 2002. ISSN 0276-7783. URL https://www.jstor.org/stable/4132319. Publisher: Management Information Systems Research Center, University of Minnesota.

Alan F. Westin. Legal Safeguards to Insure Privacy in a Computer Society. *Communications of the ACM*, 10(9):533–537, 1967a.

Alan F. Westin. *Privacy and Freedom*. 1967b. ISBN 978-1-935439-97-4.

Alan F. Westin and Harris Louis & Associates. Equifax-Harris Consumer Privacy Survey. Technical report, 1996. Conducted for Equifax Inc. 1,005 adults of the U.S. public.

Robin Whittemore and Kathleen Knafl. The integrative review: updated methodology. *Journal of Advanced Nursing*, 52(5):546–553, 2005. ISSN 1365-2648. doi: 10.1111/j.1365-2648.2005.03621.x.

Claes Wohlin. Guidelines for snowballing in systematic literature studies and a replication in software engineering. In *International Conference on Evaluation and Assessment in Software Engineering (EASE)*. ACM, 2014. URL http://urn.kb.se/resolve?urn=urn:nbn:se:bth-6463.

J. Patrick Woolley, Emily Kirby, Josh Leslie, Francis Jeanson, Moran N. Cabili, Gregory Rushton, James G. Hazard, Vagelis Ladas, Colin D. Veal, Spencer J. Gibson, Anne-Marie Tassé, Stephanie O. M. Dyke, Clara Gaff, Adrian Thorogood, Bartha Maria Knoppers, John Wilbanks, and Anthony J. Brookes. Responsible sharing of biomedical data and biospecimens via the "Automatable Discovery and Access Matrix" (ADA-M). *npj Genomic Medicine*, 3(1):1–6, 2018. ISSN 2056-7944. doi: 10.1038/s41525-018-0057-4.

Dirk A. Zetzsche, Ross P. Buckley, Douglas W. Arner, and Linus Fohr. The ICO Gold Rush: It's a Scam, It's a Bubble, It's a Super Challenge for Regulators. *Harvard International Law Journal*, 60 (2):267–315, 2019.

Jun Zhao, Reuben Binns, Max Van Kleek, and Nigel Shadbolt. Privacy Languages: Are we there yet to enable user controls? In *Proceedings of the 25th International Conference Companion on World Wide Web*, WWW '16 Companion, pages 799–806. International World Wide Web Conferences Steering Committee, 2016. ISBN 978-1-4503-4144-8. doi: 10.1145/2872518.2890590.