PhD-FSTM-2021-032
The Faculty of Sciences, Technology and Medicine

Alma Mater Studiorum – Università Di Bologna

DISSERTATION

Defence held on 28/05/2021 in Bologna

to obtain the degree of

DOCTEUR DE L'UNIVERSITÉ DU LUXEMBOURG

EN INFORMATIQUE

AND

DOTTORE DI RICERCA

IN "LAW, SCIENCE AND TECHNOLOGY"

by

Valentina LEONE

Born on 13 July 1990 in Cuorgnè (Italy)

"LEGAL KNOWLEDGE EXTRACTION IN THE DATA PROTECTION DOMAIN BASED ON ONTOLOGY DESIGN PATTERNS"

Dissertation defence committee

Dr Martin THEOBALD, dissertation supervisor
Professor, Université du Luxembourg

Dr Luigi DI CARO, dissertation co-supervisor
Professor, Université du Luxembourg

Dr Danièle BOURCIER, Chair
Professor, Université Panthéon-Assas

Dr Sabrina KIRRANE, Vice-Chair
Assistant Professor, Vienna University of Economics and Business

Dr Armando STELLATO
Assistant Professor, University of Rome "Tor Vergata"

Alma Mater Studiorum – Università di Bologna
in cotutela con University of Luxembourg

DOTTORATO DI RICERCA IN

LAW, SCIENCE AND TECHNOLOGY

Ciclo 33

Settore Concorsuale: 01/B1 - INFORMATICA

Settore Scientifico Disciplinare: INF/01 - INFORMATICA

LEGAL KNOWLEDGE EXTRACTION IN THE DATA PROTECTION DOMAIN BASED ON ONTOLOGY DESIGN PATTERNS

Presentata da: Valentina Leone

Coordinatore Dottorato

Prof.ssa Monica Palmirani

Supervisore

Prof. Luigi Di Caro

Supervisore

Prof. Martin Theobald

Esame finale anno 2021

Abstract

In the European Union, the entry into force of the General Data Protection Regulation (GDPR) has brought the domain of data protection to the fore-front, encouraging the research in knowledge representation and natural language processing (NLP).

On the one hand, several ontologies adopted Semantic Web standards to provide a formal representation of the data protection framework set by the GDPR. On the other hand, different NLP techniques have been utilised to implement services addressed to individuals, for helping them in understanding privacy policies, which are notoriously difficult to read.

Few efforts have been devoted to the mapping of the information extracted from privacy policies to the conceptual representations provided by the existing ontologies modelling the data protection framework.

In the first part of the thesis, I propose and put in the context of the Semantic Web a comparative analysis of existing ontologies that have been developed to model different legal fields.

In the second part of the thesis, I focus on the data protection domain and I present a methodology that aims to fill the gap between the multitude of ontologies released to model the data protection framework and the disparate approaches proposed to automatically process the text of privacy policies. The methodology relies on the notion of Ontology Design Pattern (ODP), i.e. a modelling solution to solve a recurrent ontology design problem. Implementing a pipeline that exploits existing vocabularies and different NLP techniques, I show how the information disclosed in privacy policies could be extracted and modelled through some existing ODPs. The benefit of such an approach is the provision of a methodology for processing privacy policies texts that overlooks the different ontological models. Instead, it uses ODPs as a semantic middle-layer of processing that different ontological models could refine and extend according to their own ontological commitments.

Acknowledgements

The work that led to the writing of this thesis would not have been possible without the invaluable support of many people to whom I extend my thanks.

First of all, I would like to thank my supervisor, Prof. Luigi Di Caro, for his great guidance during my PhD programme and for the trust he has shown involving me in different academic activities. The enthusiasm and the positive attitude in addressing his work are a source of great inspiration for me.

I am also deeply grateful to my co-supervisor, Prof. Martin Theobald. His useful suggestions on the technologies to be used in my research became part of this thesis. Moreover, the classes he teaches at the University of Luxembourg have broaden by knowledge about big data analytics techniques.

Thank you to the research group led by Prof. Guido Boella at the Computer Science Department of the University of Turin. In particular, I really appreciated the work made by Lliho Humphreys and Ilaria Amantea, who supported the evaluation of the experimental part of the thesis. Over the years, Lliho also gave me useful feedbacks for improving my presentation skills.

Another special acknowledgement goes to Serena Villata, for her guidance in the first year of the PhD programme. Working with her was an enriching experience for me.

Thank you to my colleagues of the "Law, Science and Technology" programme, with whom I shared the last three years: Giorgia Bincoletto, Chantal Bomprezzi, Federico Galli and Salvatore Sapienza. With Giorgia and Salvatore, in particular, I shared great moments both inside and outside the academic life.

Thank you to my long-life friends: Alice, Stefania, Cristina, Gaia, Giorgia, Marianna, Claudio and Francesco. Their constant presence has been a great help for me and their friendship is a precious gift to me.

Finally, thank you to my family, to whom I wish to dedicate this thesis. Thank you to my lovely grandma, my cheering brother Andrea and my beloved niece Arianna. Thank you to my parents for their unconditional support, because it does not matter how far I am from home, I know that they are always there for me.

Contents

List of Acronyms

CP Content Pattern

DPV Data Privacy Vocabulary

DPVCG Data Privacy Vocabularies and Controls Community Group

GDPR General Data Protection Regulation

NLP Natural Language Processing

ODP Ontology Design Pattern

OIE Open Information Extraction

OPP-115 Online Privacy Policies - set of 115

PoS Part of Speech

SRL Semantic Role Labelling

1 Introduction

1.1 Legal and Technological Context of the Work

The formalisation of legal knowledge into machine-readable formats has been a challenging task since the 1990s. The two decades between 1990 and 2010 were characterised by the first theoretical approaches to the formalisation of legal knowledge, that resulted in the proposal of core ontological models, like FOLaw (Functional Ontology of Law) [189], LRI-Core [27], CLO-Core (Core Legal Ontology) [83] and LKIF-Core (Legal Knowledge Interchange Format) [100]. By contrast, the second decade of this century saw a change of focus in the objectives pursued through the modelling of legal knowledge. Indeed, the efforts in the field of knowledge representation moved towards the modelling of specific legal sub-fields, as evidence of a greater awareness of the specificity which characterises each of them.

This change of focus, however, should also be put into a technological context that saw the consolidation of the Semantic Web as a *Web of Data* [16], i.e. a virtual environment for sharing knowledge. Legacy ontological languages like ONTOLINGUA [85] and DAML+OIL [101] were gradually left behind and replaced by the W3C standards for knowledge representation, like RDFS and OWL. The effort to promote the economy of information and the reuse of knowledge on the Semantic Web was also witnessed by the proposal of small, reusable ontologies, called Ontology Design Patterns (ODPs), formalised through Semantic Web standards and intended as building blocks in the process of ontology engineering [80]. As noticed by Casanovas et. al. [32], in such a technological context, "*scalability, reusability, and end user-centered approaches where taken into account to model specific legal domains*".

Among the legal fields, for two years now, the entry into force of the General Data Protection Regulation (GDPR) on May 25$^{\text{th}}$ 2018 [1] has brought the domain of the data protection to the forefront. The Regulation sets the grounds for a lawful processing of personal data and new principles for the data protection practice. Moreover, the Regulation empowers individuals with new rights and sets new obligations for different stakeholders.

The need for the involved actors to comply with the new requirements set by the GDPR encouraged the modelling of several ontologies to represent the data protection field and support approaches to automatic compliance checking with the Regulation. The Data Protection Ontology [15], GDPRtEXT [153] and PrOnto [150] are just few examples of this effort. Each of them adopts its own ontological commitment, i.e. its own perspective about the data protection domain, according to the specific modelling requirements they address. These different perspectives bring to ontological representations that are characterised by some distinctive representational choices, but that share some similarities in the way they model the knowledge related to the field of interest. If those similarities are not modelled adopting shared conceptual solutions, then the risk is to release redundant representations of knowledge, clashing with the principle of economy of information promoted by the Semantic Web.

In parallel with these efforts to conceptualise the data protection domain, several approaches started from the raw text of the privacy policies and applied Natural Language Processing (NLP) techniques mainly for: *(i)* classifying the text of privacy policies' paragraphs according to some annotation scheme, *(ii)* developing models able to provide question-answering services for queries formulated on the text of the privacy policies, *(iii)* summarising the text of those documents by selecting the relevant information in them. These approaches have often been used for the development of user-friendly systems that assist individuals in reading privacy policies [93, 199]. Indeed, those documents are notoriously long and often adopt a vague language that could hinder the understanding of their content.

By contrast, few efforts have been devoted to the mapping of the information extracted from the privacy policies on the conceptual models provided by existing legal ontologies. Moreover, even when the proposed NLP techniques are based on some annotation scheme for labelling paragraphs according to their content, this scheme is not organised in a semantic shared structure. Consequently, the outcomes of those approaches are hardly reusable outside the context of the project for which they were implemented.

A fully automated approach that extracts information from privacy policies to directly map it on some formal ontological model is hardly possible, considering the aforementioned critical aspects that affect privacy policies. Indeed, if their reading generates issues for human understanding, than even more difficulties are expected for automated approaches applied to the text of privacy policies. However, NLP techniques applied to the text of those documents should make more effort to associate their results with a semantic superstructure that could ease their reuse and integration in existing ontological models. This effort is necessary to populate ontologies with the facts of the real world, enabling the application of automatic reasoning systems on real scenarios.

1.2 Research question

The starting point for this thesis work is the definition of Ontology Design Pattern (ODP), described as a "*modelling solution to solve a recurrent ontology design problem*" [82]. Considering this definition, the main assumption of this thesis is that if an ODP should represent a recurrent ontology design problem, then evidence of this recurrence should be retrieved in the texts belonging to the domain of interest modelled by the pattern.

Considering the highlighted gap between the multitude of ontologies released to model the data protection field and the disparate approaches proposed to automatically process the text of privacy policies, this thesis investigates the following research question:

> *Is it possible to detect, within the privacy policies text, those informative scenarios for which a modelling solution provided by an ontology design pattern already exists?*

From the point of view of the research question and considering the definition of ODP, an informative scenario should be intended as an ontology design problem that needs to be formalised and whose presence is detected within the text of privacy policies. By applying NLP techniques to the text of privacy policies, this research work investigates the possibility to map the information they describe on some ODPs that are suitable to represent the data protection field. The benefits of such an approach should be twofold: *(i)* the output of the application of NLP techniques to the text of the privacy policies could converge towards a standardised conceptual model that enriches the extracted information with some semantic structure, *(ii)* the adoption of a shared representation of information, based on the conceptual models provided by ODPs, might represent a semantic middle-layer that acts as a glue between the unstructured information disclosed in privacy policies and the multitude of possibilities in representing this information according to the conceptual models provided by higher-level legal ontologies.

1.3 Methodology

The thesis acknowledges the importance of promoting the reuse of existing ontological models, providing semantic shared representations of legal knowledge. Consequently, the starting point of the research work is an analysis of existing ODPs that could be of interest for modelling legal knowledge in the data protection field. This analysis identified an ODP that has been proposed for representing the information described in privacy policies. Specifically, the ODP provides a modelling solution for those scenarios concerning the representation of a processing activity performed on personal data.

The modelling solution provided by the ontology pattern is used to guide the detection, in privacy policies, of sentences that describe data processing scenarios, mapping the information in the sentences on the most appropriate classes of the pattern. The detection of those sentences is made by an automatic system that processes the the text of privacy policies relying on open information extraction techniques and text similarity measures. Aiming at reusing other existing Semantic Web oriented knowledge sources, the system also takes advantage of a domain-specific vocabulary, named Data Privacy Vocabulary, proposing further mappings of sentence excerpts with the concepts modelled in it. The performance of the system as been evaluated by two legal experts who expressed their judgement about the precision of the system in detecting the sentences that describe a data processing scenario. The experts' assessment showed that the system can identify those sentences with a precision of 88%. Moreover, with a precision of 71.5% can match specific parts of the sentences with an appropriate concept in the pattern. The developed approach and the achieved results also showed that at least two additional ODPs could be exploited to model the information in privacy policies.

1.4 Outline of the Thesis

The reminder of the thesis is organised as follows.

Chapter 2 introduces the field of knowledge representation. The first part of the chapter provides an analysis of the different definitions of the term *ontology* and presents some topics related to ontology engineering and the Semantic Web. The second part of the chapter presents the personal contribution provided in the analysis of the existing legal ontologies, presenting a study that compares the legal ontologies based on feature-based analysis. Finally, the chapter describes an applicative experience of this analysis in the development of a system that finds existing legal ontologies based on the modelling requirements expressed by the user.

Chapter 3 introduces the data protection domain. The first part of the chapter provides an overview on the reform of the data protection domain undertaken by the European Union, then focusing on the explanation of the core principles of the GDPR. The second part of the chapter discusses some issues related to the adoption of written privacy policies as a mean to disclose data practices to individuals.

Chapter 4 provides, in the first part, a summary of the tasks that can be addressed by applying different NLP techniques on unstructured text, presenting the solutions provided by state-of-the-art systems in addressing those tasks. The second part of the chapter describes the steps that were implemented for investigating the research question, presenting the resources adopted to perform the experiments and explaining the methodological choices that resulted in the implementation of the system that is proposed for extracting recurrent scenarios from privacy policies.

Chapter 5 presents the evaluation of the proposed system, that relied on the manual assessment of two legal experts. After discussing the design of the annotation task assigned to the experts, the results of the evaluation are discussed.

Chapter 6 provides an analysis of some works that are related to the topics covered by this thesis. The first part of the chapter presents different works that are based on ODPs and the automated approaches to the processing of texts in legal fields other than the data protection one. The second part of the chapter focuses on related works in the data protection field, trying to highlight similarities and differences of the existing approaches with the work presented in this thesis.

Chapter 7 ends the thesis providing some final remarks and shaping future directions of research.

1.5 Publications

This research work builds upon the papers co-authored over the past three years. Parts of this thesis have appeared in the following publications:

– *Leone, V., Di Caro, L., Villata, S. (2020).* ***Taking stock of legal ontologies: a feature-based comparative analysis****. Artificial Intelligence and Law, vol.28(2), pp. 207-235.*
  This paper presents a comparison of the most recent ontologies that were released to model different legal domains. The paper develops a feature-based analysis for comparing ontologies from different points of view.

Specifically, the ontologies are compared with respect to: *(i)* the practical information that the ontologies disclose and that could be proved useful to enhance their reusability, i.e. their version number or licence under which their are released, *(ii)* information about the methodological and technological choices followed to build the ontology, *(iii)* the modelling choices for representing the knowledge they refer to. This work has been included and partially enlarged in Chapter 2 of the thesis;

- *Leone, V., Di Caro, L., Villata, S. (2018).* ***Legal Ontologies and How to Choose Them: the InvestigatiOnt Tool****. In: van Erp M., Atre M., Lopez V. et al. International Semantic Web Conference (P&D/Industry/BlueSky). CEUR Workshop Proceedings, vol. 2180.*
  This paper presents a demo system whose conceptual implementation grounds in the feature-based comparative analysis developed and presented in the paper above. The system helps users in finding an existing ontology that corresponds to the modelling requirements that they formulate by answering a set of questions asked by the system. This work is presented in Chapter 2;

- *Leone, V., Di Caro, L. (2019).* ***Frequent Use Cases Extraction from Legal Texts in the Data Protection Domain****. In: Araszkiewicz A., Rodríguez-Doncel V. (eds) Legal Knowledge and Information Systems. Frontiers in Artificial Intelligence and Applications. vol. 322, pp. 193-198.*
  The paper presents an application of an existing tool for performing open information extraction, i.e. ClausIE, to extract from sentences *(subject, verb, object)* triples. The results were used to provide an insight of the possibility of relying on those triples for identifying recurrent information from text and mapping it on existing ODPs. This work was the starting point of the implementative part of the thesis and it is included, in an extended version, in Chapter 4;

- *Leone V., Di Caro L. (2020).* ***The Role of Vocabulary Mediation to Discover and Represent Relevant Information in Privacy Policies****. 33 International Conference on Legal Knowledge and Information Systems (JURIX 2020).*
  This paper describes a system that relies on a domain-specific vocabulary, i.e. the aforementioned Data Privacy Vocabulary, to detect mentions to personal data types and purposes of the data processing in privacy policies, based on the concepts modelled in the vocabulary. The system also relies on a general-purpose computational lexicon, i.e. BabelNet, to find lexical variants of the terms modelled in the vocabulary. The cor-

respondences between relevant parts of the text and the concepts in the vocabulary are established on the basis of the cosine similarity computed between the sentence excerpts in privacy policies and the concepts modelled in the vocabulary. This work is presented in Chapter 4.

2 | Knowledge Representation: Opportunities and Challenges

This chapter presents the technological background of the thesis. The first part of the chapter provides an overview of the different interpretations of the term *ontology* and presents the main aspects involved in the ontology engineering process. The challenges of knowledge representation are, then, put into the context of Semantic Web that promotes the interoperability and sharing of knowledge through the provision of standards for the representation of knowledge. In this context, the definition of ODPs is presented and discussed.

The second part of the Chapter focuses on the contribution of this thesis to the analysis of the state of the art in the field of legal knowledge representation. An analysis of existing legal ontologies, developed at three levels of comparison, is presented by providing summary tables that help the comparison of the resources.

The last section of the Chapter presents an applicative outcome of the performed analysis, that resulted in the development of a Web application that helps users interested in the reuse of existing legal ontologies selecting the one that best suits their requirements.

2.1 Vocabularies and Ontologies: Shaping Similarities and Differences at Different levels

2.1.1 Ontology: One Term, Many Artefacts. Classification of Ontologies According to their Level of Formalism

The term "ontology", initially coined within the philosophical field, has been adopted by the Artificial Intelligence community since the 1990's. In its philosophical understanding, this term refers to the discipline that studies reality, i.e. *what there is* as part of the real world, how it is organised and how it could be described [182].

In the Artificial Intelligence field, many definitions of the term "ontology" have been proposed and discussed [88, 173] without reaching an agreement in defining what an ontology is. However, the nowadays common understanding of the term is that of a machine-readable artefact used to formally represent a certain domain of discourse. The representation offered by an ontology aims to make explicit both the vocabulary of terms used to describe such a domain and the set of explicit assumptions underling the intended meaning of each term [87]. The vocabulary terms are called *classes* of the ontology and they represent the relevant concepts in the domain of discourse. Specifically, an ontological class is an abstraction of the set of real objects (called *instances*) corresponding to the conceptualisation represented by the class itself.

Other typical constituents of the ontologies are the relationships that describe how concepts interrelate with each other. An example is represented by the taxonomic relationship which is used inside ontologies to organise concepts hierarchically, determining which concepts are more specific than others (the former being called *subclasses* and the latter being called *superclasses*). More relationships can be then defined depending on the domain of discourse taken into account. Additionally, ontologies can specify properties of concepts, restrictions on the values that properties can assume and other constraints expressed using some formal logic language.

The types of constituents adopted by an ontology determine its *level of formalism*. In particular, the more components are included, the higher is the level. Therefore, the term "ontology" is actually used in the literature to denote a wide set of artefacts that differ in their level of formalism. Specifically, a well-known analysis proposed by D. McGuinnes [126] identifies a spectrum of nine types of artefacts, ordered by increasing level of formalism. *Controlled vocabularies* and *glossaries* rank at the bottom of this spectrum. While controlled vocabularies are simple lists of terms, glossaries also include, for each term, a description of its meaning in natural language. *Thesauri* rank one step above in the spectrum, as they organise the list of terms according to some semantic relation. In particular, they use the synonymy relation to group terms that express the same meaning. Thesauri do not offer the possibility to represent explicit hierarchies of concepts, but they allow the specification of broader and narrower terms. By contrast, *informal taxonomies* order concepts with respect to the taxonomic relationship, but the membership of an instance to a class is granted only for its direct superclass, while the membership to the inherited superclasses is not a rigorous requirement. *Formal taxonomies*, instead, define strict hierarchies of concepts such that, if an instance belongs to a class, then it is possible to infer that this instance also belongs to all the inherited superclasses. Moving up in the spectrum of ontologies, frames define a class not only with respect to the position it holds in the hierarchical structure of concepts, but also with respect to its properties. These properties are both those directly owned by the class and those that are inherited by its superclasses. In a higher level of formalisms, an ontology can then express *restrictions on values* that these properties can assume (e.g., a data type restriction or a domain restriction [173]). Moving up to the uppermost part of the spectrum of ontologies and increasing further the level of expressiveness, it is possible to specify more constraints by means of some formal logical language. This constraints can be expressed over classes or relationships. Using constraints over classes it is possible to specify, for instance, whether a class is the result of the intersection of two other classes or whether a class is disjoint from another one. Instead, constraints over relationships allow the specification, for instance, of inverse or transitive relationships.

The fine-grained classification of ontologies provided by McGuinnes is usually reduced to a coarser classification that identifies only two categories of artefacts, i.e. lightweight ontologies and heavyweight ontologies. The difference between these two macro-groups can be understood analysing the role that the natural language plays in each group of artefacts. Lightweight ontologies, also called *linguistic/terminological ontologies* [168] or *vocabularies* (the latter naming is particularly used in the Semantic Web context, discussed in Section 2.2) usually include artefacts ranging from controlled vocabularies to informal taxonomies. Natural language is their primary focus since they aim to overcome some of its intrinsic characteristics as, for instance, the polisemy of words and their consequent ambiguity. To achieve this goal, they list the relevant terms in the domain of discourse providing a normalised and lightly structured set of lexical terms.

By contrast, formal taxonomies and all the artefacts at the higher levels of the spectrum defined by McGuinnes are usually referred to as *heavyweight ontologies*. They transcend the dependence from natural language and terms are just used as symbols in some formal logic language to define the concepts, properties, relationships and constraints on them. The use of formal logic allows the overcoming of some of the criticisms of natural language, providing unambiguous descriptions that limit the interpretation of the meaning associated to the constituents elements of an ontology [97].

2.1.2 Ontologies in Scope

As discussed in Section 2.1.1, an ontology aims to provide a formal representation of a domain of interest. When analysing the proposed representation, the *ontological commitment* of the ontology is a fundamental aspect to consider. It refers to the set of choices made to select the facets of the reality that were judged as relevant in the domain of interest and, consequently, worth to be represented within the ontology. A completely truthful and detailed replication of the entities that belong to reality is in any way impossible to achieve, because the only accurate representation of an entity is the entity itself [48]. Accordingly, every representation of knowledge is inevitably imperfect and it is only an approximation of reality. Therefore, the set of choices that form the ontological commitment to which the ontology adheres is an essential and necessary component of every knowledge representation. It could metaphorically be compared to the viewpoint from which the concrete and abstract entities of reality are observed. Depending on the chosen viewpoint, some aspects of their faces will be revealed, while others will be hidden. Consequently, different viewpoints determine the possibility of having a multitude of representations referring to the same domain of interest. Committing to an ontology means agreeing with its viewpoint, acknowledging that it properly represents the reality of its domain of interest [96].

The broadness of the domain encompassed by the ontology, i.e. its scope, is another fundamental aspect to look at. The scope of ontologies depends both on the nature of the entities that are represented and on the level of agreement that the ontology is supposed to reach among its adopters (i.e. the users committing to the ontology). Considering these two factors, ontologies are usually classified in five different groups, sorted by increasing scope [137, 168]. *Application ontologies* are developed, within a certain domain of interest, to meet the specific purpose of an application and their scope is specified through testable use cases [121]. These types of ontologies are not supposed to reach a high level of agreement between user. Instead, they reflect the single viewpoint of the developer that is testing the use case, or the user that commissioned it. *Domain ontologies* model the knowledge related to a specific domain of interest, catching the viewpoint of a group of users in their way of describing the entities belonging to the domain and the relationships that link them. When the viewpoints of different groups of users are combined to represent the central concepts of the domain of interest, the resulting artefact is called *core ontology*. Domain independence is reached by *general purpose ontologies* that represent generic knowledge (e.g., units of measurement or temporal relations) useful to link more specific concepts within domain ontologies. Finally, *foun-*

dational ontologies (also called *top level ontologies* or *upper ontologies*) achieve the widest scope by defining the most general concepts (e.g., objects, events and processes), shared among different domains and areas of interest. The domain independence of general purpose and foundational ontologies should thus reach a high level of acceptance among users, up to an ideally word-wide commitment expected for foundational ontology.

2.1.3 Approaches to Ontology Engineering

Ontology engineering is a complex and multifaceted process that involves a multitude of activities and practices that should drive the development of the ontology. Regardless of the specificity of the different available methodologies, there are certain activities and criteria that should guide the definition of any methodology for ontology engineering. The main steps of an ontology engineering methodology should include: the analysis of the domain of interest to elicit the core terminology, the conceptualisation of the terminology into a language-independent level of abstraction, the implementation of the conceptualisation in some ontological language, the evaluation of the proposed model and the population of the ontology with the instances that represent facts of the modelled domain [178]. Other good practices that increase the quality and the replicability of a methodology for ontology engineering concern the provision of a good documentation of the involved activities and methods, the grounding in an existing and consolidated methodology, the orientation to support the interoperability and the disposition of strategies for the maintenance of the ontology to address the changes of the domain over time [172]. A methodology for ontology engineering should also provide guidelines for a collaborative development of the ontology. In the life science field, the Open Biomedical Ontologies (OBO) Foundry was a good example of how the collaborative creation of ontologies could be achieved setting rules for the formulation of relational assertions, for the adoption of naming conventions and for the convergence to an agreed term when multiple possibilities from different ontologies are provided [181].

The most famous methodology for ontology development is “Ontology development 101” [145], having more than 6000 citations on Google Scholar. The methodology provides several steps for developing an ontology using the Protégé[1] tool, providing practical suggestions for avoiding common pitfalls. This methodology recommends the use of competency questions to determine the scope of the ontology and to test the achievement of the requirements that the ontology should fulfil once the development ends.

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

NeOn [185] provides a flexible workflow for ontology development. The proposed methodology does not prescribe a set of steps to follow in linear order, instead it identifies a set of nine scenarios that could occur during the ontology development and it provides a set of processes and activities that should be accomplished to handle each scenario. Some of the core principles of the methodology are the reuse of existing ontological and non-ontological resources for the development of an ontology as well as the promotion of a collaborative effort between ontology practitioners and developers.

Approaches to ontology engineering have frequently been inspired by the agile methodology of software engineering, that is based on an iterative and incremental development of the software aimed to integrate modification and change at every iteration, thus minimising the risk of failure of the overall project. Methontology [73] was one of the first ontology engineering methodologies to adopt the iterative approach. It identifies three main processes in ontology development, i.e. the management process, the development process and the support process, each of them containing some specific activities. UPON Lite [49] is a lightweight methodology for the rapid prototyping of ontology and consists of six steps. The aim of UPON Lite is to leave room to domain experts in the development of the ontology, limiting the intervention of the ontology engineers to only delivering the formal ontology. DILIGENT methodology [160] also supports and evolutionary lifecycle for ontologies focusing on the collaborative efforts of several stakeholders.

In addition to these methodologies that have embraced the general principles of agile software engineering, other approaches have adopted the practices of specific methodologies in agile programming, such as the extreme programming [23] or SCRUM [2] development, to formulate their methodologies, still pursuing the goal of a rapid and incremental development of ontologies.

2.1.4 Methodologies to Evaluate Ontologies

The previous paragraphs should have highlighted the complexity underlying the study of ontologies, that require a conceptual modelling of reality, an applicative analysis of the purpose to be fulfilled by the provided conceptualisation and, finally, methodological and representational choices to transform such a conceptualisation into a concrete machine-readable artefact. A methodology that aims for the evaluation of an ontology must necessarily take into account the coexistence of these factors, determining: *(i)* one or more *aspects* that are to be evaluated in the ontology, *(ii)* the assessment *criteria* that should guide the evaluation of each of the selected aspects and *(iii)* the *approaches* used to provide a measurable and comparable values to the criteria.

In the literature, several studies [30, 81, 106, 194] have suggested a multilayered evaluation that accounts for the different aspects contributing to the final representation of an ontology. The syntactic, semantic and functional layers are the factors commonly considered when evaluating an ontology. The syntactic layer evaluates an ontology with respect to the specific knowledge representation formalism that is adopted to write the ontology, the semantic layer assesses the meaning associated to the elements of the ontologies while the functional level examines the ontology according to the specific use that should be made of it when integrated in a complex system. Moreover, the graph-like shape of the ontologies allows a further evaluation developed at the structural layer.

The aforementioned levels of evaluation focus on the internal features of an ontology, i.e. how it is organised and represented [106]. However, ontologies can also be evaluated with respect to some external aspects like the social role that they play, concerning the leverage that ontologies have on a community of users, and their usability profile, considering how the metadata associated with the ontologies promote their understanding by interested users.

Other studies related to the evaluation of ontologies have identified different aspects to consider in the evaluation, decoupling it from the analysis of the internal and external characteristics of ontologies. In [99], the authors propose two complementary perspectives for evaluating an ontology, i.e. the quality and the correctness. The former focuses on the formal representation that the ontology provides for the domain of interest and evaluates how this representation promotes an efficient reuse of the ontology. The latter evaluates the ontology with respect to the reality, considering how much the approximate representation provided by the ontology (see Section 2.1.2) deviates from the reality itself. In [86, 106], the evaluation of an ontology should refer to its stages of development, thus distinguishing between an evaluation useful to asses the design phase, and an evaluation useful to evaluate the implementation phase in the ontology development pipeline.

As anticipated at the beginning of this section, evaluating the different aspects concerning an ontology requires the definition of specific criteria, i.e. leading parameters for the assessment of each of those aspects [194]. Even if a set of standard criteria does not exist, it is possible to identify some of those criteria that are commonly used for developing a layered evaluation of ontologies. The assessment of the syntactic layer should be steered by the lawfulness and richness criteria[30, 106]. The former assesses the extent to which the rules of the adopted ontology language have been complied when writing the ontology, while the latter considers how much the expressive power of the ontology language has been exploited (suggesting the distinction, discussed in Section 2.1.1, between ontologies, that usually include axioms in the definition of their concepts, and vocabularies that do no). The evaluation of the structural layer relies on some criteria traditionally used for describing the graph structure, like the depth and breadth of the hierarchies and their density [81, 106]. The assessment of the semantic layer of an ontology is driven by criteria like consistency and clarity(also called coherence) [30, 81, 86, 106, 194]. Consistency aims to avoid contradictions, asking for uniformity and harmony in the definitions provided for the different terms of an ontology. The consistency criterion concerns both an informal understanding that refers to the descriptions in natural language provided in the documentation of the ontology, and a formal understanding that asks for logically consistent axioms. By contrast, clarity concerns the understandably of an ontology and assesses whether the semantics of the terms encoded within an ontology is easily intelligible [137]. Finally, accuracy[30, 106, 194], comprehensiveness [30, 194], relevance [30, 81] and adaptability [86, 194] are the prevailing criteria used to evaluate the functional layer of an ontology. The accuracy criterion assesses the correctness of the information represented by the ontology with respect to the real word, comprehensiveness evaluates whether the ontology properly covers the domain of interest, relevance measures the suitability of the ontology in being able to satisfy the requirements formulated by the users and adaptability refers the ability of the ontology to be specialised, without a retreat of the already existing definitions.

The criteria used to guide the evaluation of the semantic and functional layers are all applicable for assessing the design stage in the development of an ontology [54]. Among these criteria, consistency, comprehensiveness and accuracy can also be applied for evaluating the correctness of an ontology, while clarity and adaptability should lead the assessment of its quality [99]. By contrast, the evaluation of the implementation stage of ontology development calls for the introduction of new criteria, like the computational efficiency, that assesses how much the ontology eases the processing of automatic reasoning, and practical usefulness, that assesses the number of practical problems to which the ontology applies [54].

OntoClean [89] is another well known methodology for evaluating ontologies. It is based on four criteria (i.e., rigidity, identity, unity and dependency) that are derived from Philosophy and that are used to validate the taxonomic structures within ontologies. The assessment of each of those criteria allows the detection in the taxonomy of problematic area that may need to be reviewed.

The criteria used to evaluate the external aspects of an ontology are more variable, but some similarities can be identified. The social layer of the ontology can be evaluated using criteria such as authority, that assesses the number of ontologies that are linked to the considered one, and history that counts the number of times an ontology is accessed [30]. The usability layer, instead, can be evaluated with the recognition criterion, that estimates how well the ontology is documented, and the interfacing criterion that evaluates the availability of metadata allowing a user friendly visualisation of the ontology content [81, 106].

After the definition of a set of criteria for the evaluating the distinct aspects of an ontology, different approaches can be applied for associate to each criteria a measurable and comparable value. There are four main approaches to ontology evaluation that have been traditionally distinguished in the literature [26, 146, 137], i.e. human-based, gold standard-based, data-driven and task-based. Each approach is suitable to quantify some of the previously listed criteria. The gold standard-based approach compares the developed ontology with one or more reference ontologies that are used as a ground truth in representing the relevant concepts and their relationships in the domain of interest. This approach is suitable to evaluate the functional layer of an ontology, but it could be difficult to find another ontology that was created under the same conditions and purpose, so that a fair evaluation could be made [165]. The task-based approach evaluates the extent to which the performance of an application increases when the ontology is integrated in such an application. This approach can equally used to evaluate the functional level of an ontology, but also its semantic layer and its computational efficiency. The data-driven evaluation is similar to the gold standard-based approach, but it uses a corpora of documents that refers to and sufficiently cover the domain of interest. The information emerging from the corpus is then compared with the conceptualisation encoded in the ontology. Finally, human-based evaluation is potentially applicable at every level of ontology assessment. However, it is indicated especially in the evaluation of the functional level of an ontology because a human expert owns that expertise and background knowledge that can hardly be captured in an ontological representation, but that contribute to the understanding of reality in a certain domain of interest [81]. Human intervention is also essential in the OntoClean methodology that requires experts familiar with the above mentioned criteria to express an assessment for each of them.

2.2 Vocabularies and Ontologies in the Semantic Web

2.2.1 The Semantic Web Vision and the Role of Vocabularies and Ontologies in It

The Semantic Web epitomises the enrichment of the traditional World Wide Web (WWW) with a semantic super-structure that allows the content published in it to be automatically processed by machines. Before 2000, the informative content on the Web (such as HTML pages, audio files and videos) was mainly conceived for human consumption, thus requiring a considerable computational effort to be processed by a machine. This traditional understanding of the Web is often referred to as "Web of Documents" and it has been superseded by a new perspective, called "Web of Data", which was envisioned by Berners-Lee et al. [16] in 2001. In contrast with the Web of Documents, where the informative part of a content is not explicit and asks for human intervention to find it out, the Web of Data is made by non ambiguous statements that describe objects, resources and real word facts and that can be exploited by automatic applications. This is the idea that the adjective "Semantic", coupled with the term "Web", refers to: a Web where the published content is not just machine-representable, but it is made machine-understandable through an explicit encoding of its semantics.

The ultimate goal of this new vision of the Web is to ease and enhance semantic interoperability. The fulfilment of this goal is realised when the systems are able to publish and share data on this new semantic environment, query the data published by other systems, reason about those data and eventually integrate their own data with those coming from different sources in the Semantic Web, drawing semantic links among them. Consequently, the Semantic Web becomes a global network of data connected through meaningful links that generate a collection of shared knowledge, i.e., a knowledge graph [90]. The semantic interoperability is possible when the meaning of a datum and its interaction with other data is made explicit. Therefore, the systems that interface on the Web use metadata with the purpose of representing the meaning associated to the row data they publish, enabling other systems to acknowledge and understand this meaning. The metadata are organised and structured into ontologies whose role in the Semantic Web is not different from the one they play in a legacy context: they provide a shared understanding of a domain and the data that belong to it, overcoming the differences and ambiguities in terminology [8]. This ontological level builds the meaningful and readable superstructure that enriches the data, giving more emphasis to the role that ontologies play in the Semantic Web context. They are the backbone of this new vision and one of the fundamental aspects that distinguish the traditional human-centred notion of the Web with the new machine-oriented perspective.

For the objective of semantic interoperability to be real, it is necessary to create a common technological ground for the systems that interface and operate in the Semantic Web. Accordingly, the realisation of the "Web of Data" vision has been, at least at the beginning, an engineering and technological challenge rather than a scientific one [8]. The pivotal idea of this technological challenge was the agreement in associating a URI (Uniform Resource Identifier), specifically HTTP URI, to each resource (e.g. objects, facts, things) on the Web. This naming convention, and the adoption of XML (eXtensible Markup Language) as the reference standard for exchanging structured documents, underpin more advanced standards that realise the semantic layer of the Web. Specifically, RDF enables the assertion of statements about the Web resources that can be related according to terminological criteria through the SKOS standard. Moreover, through the use of RDFS and OWL, objects can be interconnected through semantic links of variable complexity. The SPARQL standard can then be used to formulate queries about those data. Those standards can be visualised through the so called "Semantic Web Layer Cake" that was initially proposed by Berners-Lee[2] and has been updated over the years based on the convergence to one or more standards for the implementation of the different levels [3]. Section 2.2.2 contains a more detailed description of each of the aforementioned standards that realise this stack of standards.

The exploitation of the URI for naming the resources on the Web and the adherence to some good practices for sharing and semantically link these resources creates a dense network of data, called Linked Data [4]. The Linked Data ecosystem is growing year after year and counted 1255 open dataset linked by 16174 relations , on March 2020[5]. In the last years, the volume and the increasing heterogeneity of the data shared on the Web is reshaping the role that ontologies play in this context. The formal semantic provided by the ontology

[2]The first version of the Semantic Web Layer Cake buy Berners-Lee is available at https://www.w3.org/2000/Talks/1206-xml2k-tbl/slide10-0.html

[3]The first updated version of the Semantic Web Layer Cake was proposed by Berners-Lee in 2006 and is available at https://www.w3.org/2006/Talks/0718-aaai-tbl/Overview.html#(14). Later and more recent versions have been discussed in the article by B. Nowack in 2009 available at http://bnode.org/blog/2009/07/08/the-semantic-web-not-a-piece-of-cake and in the article by K. Uyi Idehen in 2017 available at https://medium.com/openlink-software-blog/semantic-web-layer-cake-tweak-explained-6ba5c6ac3fab

[4]The Linked Data principles were discussed for the first time by T. Berners-Lee in https://www.w3.org/DesignIssues/LinkedData.html

[5]Data available at https://lod-cloud.net/

languages of the Semantic Web is hardly applicable when diverse data with sometimes contradicting nature exist and asks for decrease the level of formality in encoding semantic to increase, by contrast, the applicability of these ontologies to a wide set of disparate data [17]. The difficulty in reaching a satisfactory comprehensiveness of wide, largely formal ontologies explain the success, in the last years of simple specialised ontologies that offer a small set of widely applicable concepts to reach a high level of agreement in describing a set of information widely used across different domains [58]. Some examples of these widely used ontologies are the Dublin Core[6] set of metadata for describing archival information, the Friend-Of-A-Friend (FOAF) ontology[7] for describing persons, objects and their interactions and vCard[8] ontology that describes people and organisations. The comprehensiveness of these ontologies is suitable to enhance interoperability and integration over data because they offer a common semantic layer that can eventually be extended by different systems according to their meaning.

2.2.2 Representing Vocabularies and Ontologies in the Semantic Web

2.2.2.1 RDF

The Resource Description Framework[9] (RDF) is the W3C standard that specifies the abstract syntax (also called "data model") for representing and exchanging information about entities in the Web. It relies on the notion of graph to describe and identify the syntactic elements that are part of the framework. Specifically, an RDF graph is a set of *statements*, namely triples consisting of a *subject*, a *predicate* and an *object*. The subjects and the objects of such triples are the nodes of the graph, while predicates correspond to the directed labelled arcs that connect a subject to an object.

The nodes of an RDF graph represent *resources*, i.e. any entity that is part of reality. Resources are identified by an URI (Uniform Resource Identifier) or a literal value. Literals are UNICODE strings always coupled with a datatype, that enables the correct interpretation of the value that the string represents. A datatype is itself a IRI that refers to the vocabulary containing the definition for that datatype. For instance, the literal `"1"^^xs:integer` represents the resource *"1"* as a string that should be interpreted as an integer, as defined in the XML Schema[10] (represented by the `xs` namespace). Additionally, when strings

[6] https://dublincore.org/specifications/dublin-core/dcmi-terms/
[7] http://xmlns.com/foaf/spec/
[8] https://www.w3.org/TR/vcard-rdf/
[9] https://www.w3.org/TR/rdf11-concepts/
[10] https://www.w3.org/TR/xmlschema11-2/

represent natural language terms or sentences, the literal may be associated with a tag representing the language used. Literals can appear only as objects of a statement. By contrast, resources identified by a URI can appear both as a subject and object of a triple. The properties that connect a subject and an object in the RDF graph are also resources and, consequently, they are also represented by an URI. The basic RDF syntax only allow the representation of binary relations.

The graph-based view is not the only possible representation of the RDF syntax. Each statement can also be though of as a logical formula $P(x, y)$ where P is a binary predicate that relates a subject x and an object y [8]. A conjunction of binary predicates form the RDF graph.

Independently from the abstract conceptualisation of a set of RDF statements, i.e. as a graph or as a conjunction of binary predicates, the RDF triples can be represented in different machine-readable formats, called *serialisations*, that differ for their concrete syntax. Some of the most used serialisation are: RDF/XML[11] that is based on the XML syntax, the Turtle[12] serialisation that offers a more compact and human-friendly syntax and the JSON-LD[13] serialisation that is intended to support the use of Linked Data in web programming. A database storing RDF statements is called *triplestore* and usually provides interfaces useful to formulate semantic queries about the statements.

The expressive power of RDF is limited as its syntax only allows the encoding of information about individual resources. The next paragraph will present the RDF Schema that was introduced as semantic extension of the RDF data model.

2.2.2.2 RDF Schema

RDF Schema[14] (RDFS) makes it possible to formally organise the multitude of RDF statements that can be asserted about any resource associated with an URI.

The RDFS data-model allows the description of homogeneous groups (i.e. classes) of resources via `rdfs:Class`. Classes are themselves RDF resources and the members (i.e., instances) of a class are stated via the `rdf:type` property. Relationships that link two classes are instances of `rdfs:Property`, that represents the class of RDF properties. Similarly, the class `rdfs:Literal` represent the set of all literal values. Everything in the RDFS model is a resource

[11] https://www.w3.org/TR/rdf-syntax-grammar/#section-Syntax
[12] https://www.w3.org/TR/turtle/
[13] https://www.w3.org/TR/json-ld/
[14] https://www.w3.org/TR/rdf-schema/

represented by `rdfs:Resource` and all other classes (such as the aforementioned `rdfs:Class`, `rdfs:Property` and `rdfs:Literal`) are subclasses of this class.

Using RDFS, property restrictions may be stated for specifying to which classes the instances participating in a relationship P must belong. Specifically, the property `rdfs:domain` is used to assert the class to which an instance that participates as a subject in the relation P must belong. Similarly, the `rdfs:range` property specifies the class to which an instance that participates as object in P must belong to. RDFS also provides the vocabulary for the definition of hierarchical relationships, both for classes and properties, through the properties `rdfs:subClassOf` and `rdfs:subPropertyOf`, respectively. Both properties define transitive relationships.

Because the RDF and RDFS vocabularies are based on model-theoretic semantics[15], the use of those vocabulary to express knowledge allows the drawing of valid logical inferences, that generate new knowledge from the existing one.

2.2.2.3 OWL

The Ontology Web Language[16] (OWL) enables a higher level of semantic expressivity with respect to RDFS. The most recent version of OWL was released in 2012 and it is called OWL 2. It was proposed as a revision and extension of the first version of OWL, i.e. OWL 1, published in 2004. From here on, the discussion will focus on OWL 2, although many of its features are also common to OWL 1. OWL 2 is released in several versions, called *profiles*[17], that allow users to choose, according to their specific requirements, the adequate balance between expressive power and computational effort in automated reasoning.

An OWL ontology is usually made of two parts: (*i*) a terminological knowledge (T-Box) that describes a domain of interest in terms of general concepts, represented by classes, and the properties that hold among them; *(ii)* an assertional knowledge (A-Box), that expresses statements about concrete objects, i.e. instances, of the domain of interest, complying with the high-level description provided in the T-Box.

OWL inherits some of the RDFS constructs, but it also introduces a new vocabulary that increases the expressive power, still preserving the RDF triple-based representation. There are three main syntactic categories in OWL: the *entities* (e.g. classes, properties, individuals) are the building block of an on-

[15] https://www.w3.org/TR/rdf11-mt/
[16] https://www.w3.org/TR/owl2-primer/
[17] https://www.w3.org/TR/owl2-profiles/

tology and are identified by an URI, the *expressions* describe complex notions of the domain, typically, setting restrictions on the basic entities, the *axioms* assert what is true in the domain of interest.

With OWL, class expressions allow the description of new classes enumerating the individuals that belong to it , declaring some restrictions on properties or starting from existing classes. The class axioms allow the definition of hierarchical, equivalence and disjointedness relationship between classes.

Similarly to RDFS, properties can also be described according to their characteristics, specifying their range and their domain. Additionally, the property axioms may describe a property with respect to the relation of equivalence with another class, or for being its inverse, its complement or a subclass. A property can also be declared as being transitive or symmetric.

At the A-Box level, OWL provides a set of axioms for stating *assertions* (or *facts*), i.e. axioms about individuals. With these axioms, two individuals having different names can be asserted as being the same individual or, by contrast, being different. Individuals can also be related to data values through a special type of property, called data property.

The possibility to define axioms allows the application of automatic reasoners to an ontology formalised in OWL and the possibility to infer new knowledge starting from the formal description of a domain provided in the OWL syntax.

2.2.2.4 SPARQL

SPARQL[18] (SPARQL Protocol and RDF Query Language) is the standardised language for querying data expressed following the RDF semantics. To retrieve the triples of interest from a triplestore, a SPARQL query specifies the conditions that the triples must fulfil. These conditions are expressed in the query using the RDF triple pattern, substituting the subject, the predicate and/or the object of the triple, with variables.

Triplestores usually expose SPARQL endpoints, i.e interfaces on the HTTP network able to receive and process SPARQL queries. Those endpoints thus offers the possibilities of making available the triplestores over the Web providing query capabilities to external users.

2.2.2.5 SKOS

The SKOS (Simple Knowledge Organization System) standard[19] is an RDF vocabulary for representing, sharing and linking knowledge organisation systems

[18] https://www.w3.org/TR/sparql11-query/
[19] https://www.w3.org/TR/skos-primer/

through the Semantic Web. A knowledge organisation system (KOS) models the knowledge of a domain of interest, representing the underlying semantic structure according to some organisational schema (e.g. lists, taxonomies and thesauri) and providing facilities to enrich this structure with labels, definitions and relationships [201].

The basic building block of the SKOS data model is the concept (`skos:Concept`), i.e. a unit of thought that is independent from the different lexicalisations that express it in natural language. Each concept can be enriched with a set of multilingual labels used to expressed the preferred and alternative lexicalisations for a concept or for expressing hidden lexicalisation used for indexing purposes only. Similarly, the SKOS vocabulary provides elements for enrich a concept with documentary notes, like definitions, examples or scope notes. Relationships between concepts enable the representation of taxonomic or part/whole relationships as well as associative ones. A set of concepts is organised in a concept scheme (`skos:ConceptScheme`) that can be related to other concept schemes, by means of relationships that express exact, close, broad, narrow or related matches.

The advantage of the SKOS data model is to act as a glue between the formal representations provided by ontologies languages like OWL and the unstructured language-dependent Web contents. As the SKOS documentation clearly summarises: "SKOS can also be seen as a bridging technology, providing the missing link between the rigorous logical formalism of ontology languages such as OWL and the chaotic, informal and weakly-structured world of Web-based collaboration tools, as exemplified by social tagging applications."[19]. Thus, the vision of the SKOS standard reaches its full potential when it is used in synergy with the other Semantic Web standards, by filling the gap between language-dependent and abstract representations of knowledge.

2.2.3 A Solution to Enhance Knowledge Reuse: the Ontology Design Patterns

In the Semantic Web context, Ontology Design Patterns (ODPs) promote the economy of information and sharing of knowledge. As already mentioned in the introduction of this thesis (see Section 1.2), Gangemi and Presutti [82] defined ODPs as *"modelling solutions to solve recurrent ontology design problems"*. The concept is borrowed from software engineering, where design patterns help developers in finding standardised solutions to common problems in the design of a system. Following a similar rationale, ODPs have been proposed as motivated ontologies that can be used as building blocks in ontology design.

There are six different categories of ODPs: *(i) structural ODPs* aim to solve

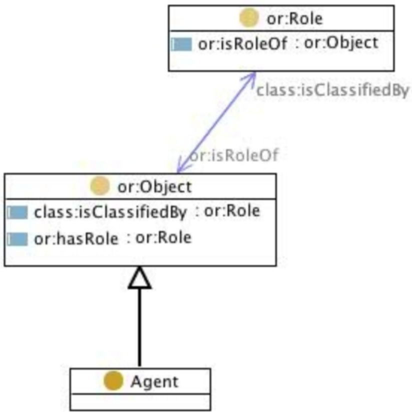


Figure 2.1: The *Agent Role* content pattern, in a graphical representation taken from the ODPs portal[21].

both expressivity issues related to the limits of the chosen ontology language and to provide architectural solutions for building the ontology, *(ii) reasoning ODPs* provides facilities to obtain some reasoning results, *(iii) correspondence ODPs* allow the mapping between ontological patterns expressed in two different formalisms, *(iv) presentation ODPs* enhance usability and readability of the ontology, *(v) lexico-syntactic ODPs* link syntactic and the ontological structures, *(vi) content ODPs* address conceptual issues, proposing classes and properties to model a specific design problem. Over the years, the Ontology Design Patterns Portal[20] [163] collected several contributions to the proposal of new ODPs, thus becoming the main reference on the Web for disclosing new ODPs (more details about the organisation of the portal will be provided in Chapter 4, when discussing the experimental part of the thesis).

Content Patterns (CPs) are of particular interest in the context of this dissertation. The design problems that a content pattern should address are characterised by two components: a domain and a use case. The same domain could encompass a multitude of use cases (i.e. possible scenarios in it) and a use case could occur in more than one domain. The identification of a use case is possible by means of the elicitation of some competency questions, that represent the queries that an expert might want to submit to a knowledge base related to her domain of interest. The set of competency questions represent the design problem that needs to be addressed, while the CP represents the solution to that problem. An example taken form the Ontology Design Pattern Portal can help in understanding how domains, competency questions and CPs

[20] http://ontologydesignpatterns.org

are related. Suppose we want to model the different roles that can be filled by staff members in a hospital (e.g. nurse, doctor, head of department). To model this scenario, the following questions are formulated: *(i) which agent does play this role (e.g. the role of doctor)?*, *(ii) what is the role that is played by that agent (e.g. a specific doctor in the staff)?*. The *Agent Role* pattern[21], shown in Figure 2.1, provides a modelling solution to this scenario relying on three classes, i.e. *Object*, *Agent*, *Role*. The *Agent* class is a subclass of *Object*, that allows the representation of both physical and social objects. The *Object* class is linked to the *Role* class by the property *isClassifiedBy* or, conversely, the *Role* class is linked to the *Object* property by the relation *isRoleOf*.

Some good practices for the proposal of a new CPs require their formulation as autonomous components. The advantage is the possibility to compose, expand, specialise and generalise those components according the specific modelling requirements, to form the target ontology. Moreover, CPs should be cognitively easy to understand and should possibly be grounded is some syntactic pattern that emerges from natural language. The formulation of the CPs should be independent from the specific ontological language adopted, but the ODPs portal usually provides their formalisation in OWL. Further discussion about the CPs in the ODPs portal will be provided in Chapter 4, in the context of the steps implemented to investigate the research question.

2.3 Vocabularies and ontologies in the legal field

After the introduction of the technological background of the work, this section presents the first contribution of the thesis. The aim is to provide an analysis of existing legal ontologies, comparing them from different points of view and analysing the extent to which the theoretical foundations of knowledge engineering, presented in the previous sections, are used in practice to model knowledge in the legal field.

2.3.1 Ontologies and vocabularies to model different legal domains

In the past years, some studies aiming at analysing and classifying legal ontologies have already been published. Casellas [33] proposed a comprehensive survey about legal ontologies spanning a fifteen-years' time range approximately, from early 90's to 2011. The features she considered in her work mainly concern the intended use of an ontology, its level of generality (core or domain) and degree of formalisation, the methodology used to build and evaluate it,

[21] http://ontologydesignpatterns.org/wiki/Submissions:AgentRole

and its availability for reuse.

Recently, de Oliveira Rodrigues et al. [50] enlarged the time-frame considered for proposing a literature review about legal ontologies published from late 90's to 2017. Their work presents different classification studies which group ontologies among different dimensions, some of them similar to those already proposed by Casellas. The new categorisation dimensions introduced by the authors concern the country and the venue where the literature about an ontology was published, its underlying legal theory, the syntactic and semantic peculiarities of legal texts that were addressed while producing the ontology (e.g., the dynamism of normative texts or the overlap of jurisdictions) and the legal subdomain it models.

If, on the one hand, the work of Casellas seems now out of date due to the lack of many recently developed ontologies, in [50] literature review it is difficult to identify the current emerging trends in the legal field due to the wide temporal interval their study focuses on. Moreover, their analysis was mainly developed on a theoretical level, relying on the scientific papers published to describe the ontologies. Features emerging from the documentation and the actual implementation of the resources, when available, seem not to have been taken into account. However, when evaluating an ontology for reuse or extension, the experts involved in the ontology building task need to consider a wide set of details. Usually, those details are not limited to the theoretical features of an ontology, but also include more practical information, e.g. the on-line availability of the ontology source file or the presence of a specific class inside the ontology.

Starting from these considerations, the classification of legal ontologies can be pushed one step further by collecting the details of their implementation and including practical information concerning their actual availability for reuse. As an ideal continuation and extension of the work of Casellas, this section proposes a comparative analysis of the legal ontologies released in the last decade, by the addition of two older ontologies which are still well known and used, i.e. Eurovoc[22] and ELI[23]. I chose to exclude from this analysis the ontologies whose source files are not available for the download, in order to enable readers to focus only on those resources that are actually available to reuse. Only two ontologies do not accomplish this requirement, i.e. ELTS [4], and PrOnto [149]. This is because they are more recent works and the eventuality that they will be released can be still considered as possible. Moreover, only the resources that model a legal domain referring to some European or

[22]https://publications.europa.eu/en/web/eu-vocabularies/th-dataset/-/resource/dataset/eurovoc

[23]https://publications.europa.eu/en/web/eu-vocabularies/model/-/resource/dataset/eli

globally applicable legal framework were taken into account. By contrast, the ontologies that focus on a national jurisdiction were excluded.

According to these selection criteria, I identified a set of seventeen ontologies belonging to five domains related to different legal fields:

1. *Policies*: it refers to the ontologies which model the permitted, mandatory and prohibited actions that can be made on a digital or material asset;

2. *Licences*: it includes the ontologies modelling the actions allowed on a resource protected by the intellectual property rights;

3. *Tenders and procurements*: this domain includes the ontologies which model the processes used by public administrations and authorities to find contractors to entrust with services or supplies;

4. *Privacy*: the ontologies model the concepts concerning the protection of personal data;

5. *Consumer Law*: it refers to the ontologies modelling the protection of consumers.

Each domain is characterised by the different sources of law it refers to and by a distinctive jargon usually reflected in the classes and properties' names of each related ontology.

In addition to the aforementioned domains, another set of four "cross-domains" ontologies is analysed. These ontologies are difficult to associate to a specific legal field because they were proposed as a more generic model for expressing deontic operators (Normative Requirements Vocabulary [79]), representing the content of legal texts in a machine-readable format (LegalRuleML [148]) and indexing documents for search (Eurovoc and European Legislation Identifier).

Figure 2.2 shows the distribution of the ontologies across the aforementioned domains. The following part of this section provides a short description of each ontology in the identified legal fields.

2.3.1.1 Policies

Open Digital Rights Language[24] (ODRL) is a language promoted by the ODRL Community Group [25] in order to model policies for digital content and media ([52]). ODRL offers a Core Vocabulary to specify the minimum set of terms

[24] https://www.w3.org/TR/odrl-vocab/
[25] https://www.w3.org/community/odrl/

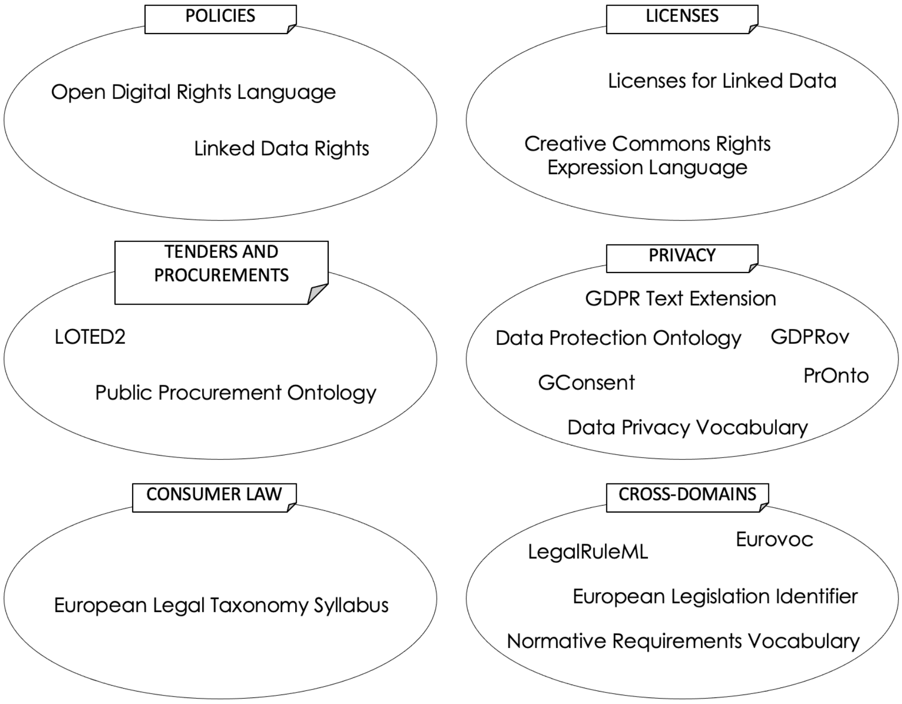


Figure 2.2: The six domains according to which the ontologies are grouped.

suitable to model different types of policies and a Common Vocabulary of general terms to model, for example, actions regulated by the obligations, permissions and prohibitions expressed in the policies. Moreover, with ODRL it is possible to associate a policy with some meta-information concerning, for example, its creator, its coverage (i.e., the jurisdiction applied upon the policy) and the versioning of the policy.

The Linked Data Rights (LDR) ontology[26] was developed by the Ontology Engineering Group[27] and it is specifically designed to model the rights which can be exercised on a Linked Data resource. LDR ontology is based on ODRL from which it extends some of the classes in order to model the conditions of use of the Linked Data resources. Specifically, LDR defines three subsets of the ODRL *Action* class in order to represent the actions permitted on a resource protected by the intellectual property rights, to use a database of Linked Data and to access a resource via the REST and SPARQL services. Moreover, it defines different types of Linked Data resources and the types of policy that can be concluded. This ontology contains also a reference to the intellectual property

[26] http://oeg-dev.dia.fi.upm.es/licensius/static/ldr/
[27] http://www.oeg-upm.net/

rights, even if they are not the main focus.

2.3.1.2 Licences

The Licence for Linked Open Data (L4LOD)[28] vocabulary uses a light ontological structure to organise the terms concerning licensing in the Web of Data. The deontic operators (permission, prohibition, obligation) are further specified in order to detail which actions can be necessarily or possibly made and avoided on Linked Open Data sources.

The Creative Commons Rights Expression Language (ccREL)[29] is the standard promoted by Creative Commons[30] (CC) to express the copyright licensing terms in a machine readable way. The ccREL ontology models all the relevant actions provided by the Creative Commons standard, distinguishing among permissions, requirements and prohibitions. All of them are further specialised by the actions which allow the sharing of a work with third parties while maintaining the copyright. Moreover, the ontology allows the specification of the legal jurisdiction which applies to the licence to be represented.

2.3.1.3 Tenders and public procurements

LOTED2[31] [59] represents the knowledge concerning the public procurements domain in the European Union, extracting the relevant lexicon of the field from a platform for public procurements' publishing. This lexicon is organised into an ontological structure legally rooted on two European Union directives about the public contracts field. Starting from these directives, LOTED2 models the roles that an agent can play in the process, the types of competition and documents used for the publication of a notice, the legal resources that regulate the field and the offers submitted for awarding a public contract.

The Public Procurement ontology[32] (PPROC) [135] models the information published in official procurement documents, focusing on the Spanish and EU law. It provides a classification of contracts according to different aspects and it allows the specification of the criteria used for the evaluation of a tender. The agents involved in a contract are expressed in the form of roles played during its execution and some hierarchies of roles are modelled. In its attempt to model the public procurements and tenders domain, PPROC makes a big effort to try to reuse information already modelled in other existing ontologies, lim-

[28] http://ns.inria.fr/l4lod/v2/l4lod_v2.html
[29] https://www.w3.org/Submission/ccREL/
[30] https://creativecommons.org/
[31] https://code.google.com/archive/p/loted2/source
[32] http://contsem.unizar.es/def/sector-publico/pproc.html

iting the introduction of new classes and properties to very specific modelling requirements.

2.3.1.4 Privacy

As mentioned in the introductory chapter of this thesis (see Section 1.1), the entry into force of the GDPR[33] in 2018 has received considerable attention from the knowledge engineering community. Since then, several ontologies have been released to model the legal framework set by the Regulation.

The Data Protection Ontology [34] [15] has been developed as part of a more complex system where it plays the role of a knowledge base for representing data protection requirements within a workflow model (e.g., a business process). The ontology models the principles related to the processing of personal data, as well as the obligations and the rights for different stakeholders involved in the processing. In particular, the ontology focuses on making explicit the relation between the rights of the data subject (i.e. the person to whom the personal data refer) and the corresponding obligations to guarantee those rights, addressed to the data controller (i.e. the entity that exercises "the overall control over the purposes and means of the processing of personal data"[35].).

The aim of GDPRtEXT[36] (GDPR text extensions) [153] is twofold. First, it aims to represent the text of the GDPR as a Linked Data resource, assigning an URI to each relevant part of the document. Second, the ontology provides more than 200 classes to represent the relevant concepts introduced by the Regulation. Those concepts refer to the categories of personal data, the agents and the activities involved in the processing of such data, the rights of the data subject and the obligations of each agent which deals with personal data.

PrOnto (Privacy Ontology) [149, 150] is another ontology that addresses the legal framework set by the GDPR to provide a model on which techniques of legal reasoning and compliance checking could be applied. In its ontological model, PrOnto makes explicit the distinction between agents and roles, the former being able to cover particular roles inside different contexts and for a limited interval of time. Moreover, PrOnto models the sequence of actions aimed at processing personal data. Besides the traditional deontic operators (i.e., permissions, prohibitions, obligations and duties), PrOnto includes concepts for modelling the compliance with and violation of an obligation. The ontology was developed with a modular approach and relies on an extensive use of existing ontology design patterns.

---

[33] The legal framework set by the GDPR will be covered in Section 3.1.2 of Chapter 3.
[34] `https://bit.ly/2uhumDv`
[35] Souce: `https://bit.ly/2PQJ5E5`
[36] `https://bit.ly/3rJTeke`

GDPRov[37] [154] starts from the acknowledgement that consent is one of the legal grounds for a lawful processing of personal data, according to the legal framework set by the GDPR. Consequently, the ontology provides the abstract model of a system for recording data processing activities and storing information about how consent for the performance of such activities was obtained, updated and eventually withdrawn. GDPRov extends the PROV-O ontology to describe provenance metadata for a planned activity and the P-Plan ontology to represent actual executions of those plans.

The concept of *consent*, as modelled by the GDPR, is also at the centre of the GConsent ontology [152]. While, on the one hand, GDPRov is more focused on modelling the provenance of consent for performing different processing activities on data flows, on the other hand, GConsent is more focused on modelling consent as an entity by itself. In the GConsent ontological commitment, the provenance is only one of the multiple aspects concerning consent. In addition to provenance, GConsent models the personal data and the purpose of the processing for which the consent was given, the state of the consent (i.e. valid or invalid) and its context (e.g. the location, the medium and the instant its was given). The ontology has been created, first, formulating a set of use-cases and the corresponding competency questions concerning the provision of consent. Second, through an iterative process, the ontology has been developed and tested based on the collected use-cases, making some adjustments when the drafted ontology was not able to address one of the identified competency questions.

The overview about the legal ontologies for the privacy domain ends with the DPV[38] [156], released in July 2019 by the W3C Data Privacy Vocabularies and Controls Community Group (DPVCG)[39]. While the other resources described in this section adopt a higher level of formalism, the DPV relies on a light-weight approach for providing a vocabulary of terms related to the field of personal data protection. The vocabulary organises the terms in several taxonomic structures, based on the specific aspects involved in the personal data handling framed by the GDPR. The DPV models, among the others, taxonomies of personal data, categories of processing operations and purposes of the processing. The top-level classes in each taxonomy are linked together through the so-called "base ontology", which has the *PersonalDataHandling* class as root concept. This vocabulary has been largely used in the experimental part of the thesis, aimed at detecting informative scenarios from the text of privacy policies. Consequently, further details and comments about this resource will be

[37] https://bit.ly/3pBf8E9

[38] https://dpvcg.github.io/dpv/

[39] https://www.w3.org/community/dpvcg/

provided in Chapter 4.

2.3.1.5 Consumer Law

The European Legal Taxonomy Syllabus (ELTS) [4] is based on the EU consumer protection law. The authors describe ELTS as a lightweight ontology, lacking of an axiomatic formalisation. This choice was made to handle the specificity of the consumer protection law at the European level as well as in each national jurisdiction in the European Union. ELTS models an ontology to represent the domain concepts of the European level and a separate ontology for each Member State to represent the concepts of their national jurisdiction. Moreover, to manage the multi-lingual landscape of the European Union, ELTS associates to each concept at the European level the corresponding lexicalisations in all the Member States languages. By contrast, the concepts belonging to an ontology at the national level are associates only with terms in the corresponding national language.

2.3.1.6 Cross-domains ontologies

Eurovoc[40] is a multilingual and multidisciplinary thesaurus managed by the Publications Office of the European Union to index the documents issued by the EU Institutions in order to ease their retrieval. The concepts are organised in 21 sectors which in turn are composed by micro-thesauri. Each sector concerns a field of competence of the EU and each concept can be associated with only one sector to avoid ambiguities (except for the sector *Geography* which allows a polihierarchy). Each concept is lexicalised by a set of terms in all the 23 languages spoken inside the EU. The terms in Eurovoc are also linked to each other through some semantic relations: beside the classical hierarchical one, also associative relations can be found among terms that are semantically related but are not on the same hierarchical structure.

LegalRuleML[41] [148, 12], is a project promoted by the OASIS LegalRuleML Technical Committee[42] which aims to develop a standard for the legal knowledge representation and exchange. To reach this goal, LegalRuleML offers a markup language which permits the harmonisation of different types of legal texts, such as norms, guidelines and policies. It provides a rich set of concepts and properties which enable the management of the complexities of a formal

[40] https://publications.europa.eu/en/web/eu-vocabularies/th-dataset/-/resource/dataset/eurovoc

[41] https://docs.oasis-open.org/legalruleml/legalruleml-core-spec/v1.0/legalruleml-core-spec-v1.0.html

[42] https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=legalruleml

representation of legal texts in a machine-readable way. Among its distinctive features, LegalRuleML provides some parameters to model the different interpretations that could be associated to a rule, to keep track of the author of a document or its fragments, to manage the temporal evolution of the norms and to take into account the defeasibility of the law.

The European Legislation Identifier (ELI) ontology[43] [76] provides a shared and uniform set of metadata for the publication of legal documents of different European Union countries to enhance interoperability among the national administrations. The ELI ontology reflects many of the basic principles of FRBR (Functional Requirements for Bibliographic Records) vocabulary[44], contextualising them into the legal field. The ELI ontology describes a legal document through the concepts of *legal resource*, *legal expression* and *format*. Since the documents issued by different EU countries could be described with different metadata according to the national jurisdiction they refer to, the ELI ontology overlooks these differences in order to represent only the common metadata of the national legal documents, providing the user the possibility to personalise and extend the set of metadata according to her needs.

The Normative Requirements Vocabulary[45] (NRV) [79] extends LegalRuleML and exploits the standard frameworks offered by the Semantic Web to represent normative requirements and rules. Differently from other existing legal ontologies, NRV is not limited to the representation of the three main deontic operators (i.e. permission, obligation and prohibition), but it specifies and organises them in a hierarchical structure according to different criteria which concern: the need for compensation, the possibility to breach or fulfil a requirement and the temporal aspects involved in their validity and compliance.

2.3.2 A three-dimensional classification of legal ontologies and vocabularies

After an overview about the fields covered by the analysed legal ontologies, this section contains a description of each feature used to classify and compare them. The overall set of features is organised in three macro-classes, called *dimensions*, according to the type of property modelled by the features they include. More specifically, the three dimensions are:

- *informational dimension*: it contains several features about the ontology disclosure and the purpose of its creation;

[43] https://bit.ly/3sk4Ucw
[44] https://sparontologies.github.io/frbr/current/frbr.html
[45] https://ns.inria.fr/nrv/v1/nrv_v1.html

- *representational dimensions* class: it refers to the methodological and technological choices made to develop the ontology, from the elicitation of knowledge to the evaluation methodology;

- *semantic dimension* class: it groups all the features concerning the way in which the ontology models the knowledge it refers to.

As mentioned before, each of these macro-classes is a set of more specific features which are detailed in Table 2.1. The following part of the section describes each feature used to classify the analysed legal ontologies.

Table 2.1: The macro-classes and the corresponding features used to classify the legal ontologies.

Macro-classes	**Features**
informational dimension	extended name legal domain purpose year current version licence updates frequency references link
representational dimension	development construction language knowledge sources for terms extraction external vocabularies references ground ontology level of structure knowledge representation formalism axioms design patterns evaluation
semantic dimension	modelling of temporal aspects adopted normative model deontic logic model

2.3.2.1 The informational dimension

The features contained in this class refer to the generic purpose for which the ontology was built together with some practical information for those who are actually interested in using the resource. Eight features belong to this class. The first information concerns the *extended name* of the ontologies. As they are often referenced by their acronyms in literature, their full name could provide to the reader a first insight of the scope of the ontology, also helping her to memorise the acronym itself. The *legal domain* feature refers to one of the six domains listed in Section 2.3.1 and it corresponds to the visual information represented in Figure2.2. This feature is further specified by *purpose* which contains a brief description of the main scope and function of the ontology inside the specified domain. Finally, the *year* feature indicates the year of the ontology's first release.

Together with this general information, some more specific features provide the reader with useful information concerning the retrieval of an ontology on the Web and its reuse. To this purpose, the *current version* feature refers to the most recent released version of the ontology, while *licence* provides the information concerning the licence under which a resource is made available for reuse. Such feature could help interested users to fairly use the ontology, respecting any limitation and constraint in its adoption. The *updates frequency* feature represents an assessment of the frequency of updates made to an ontology. Its possible values are: *low*, *medium* and *high*. and they are followed by the date of the last update. This feature is important to understand if the resource already reached a stable point and to evaluate if it is kept up-to-date according to the changes of the domain that it models. The *references* feature provides an estimate on how much an ontology is known. In particular, it corresponds to the number of references to the paper describing the ontology (and included in the bibliography of this chapter) from its publication date, as returned by Google Scholar[46]. For the resources which do not have a reference paper, it corresponds to the number of citations starting from 2012, i.e. the year the analysis proposed in the chapter starts from (see Section 2.3.1). Two research keywords were used: the first one contained the extended name of the ontology followed by the term "ontology" (except for Eurovoc, where term "thesaurus" was used as it is usually associated to this resource), while the second one contained the corresponding acronym (if available) followed again by the term "ontology". The two keywords were then linked by a disjunction operator (i.e., the OR operator). For instance, for the ELI ontology the following string was built: *"European Legislation Identifier ontology" OR "ELI ontology"*, where the quote marks were used to obtain only exact matches. Finally, the *link* feature specifies the at-present active link to the Web page containing the ontology documentation. Usually, if available, this Web page also contains the link to download the ontology source file. Table 2.2, Table 2.3 and Table 2.4 classify the ontologies presented in Section 2.3.1 according to these features.

[46]https://scholar.google.it/

	Eurovoc	**ccREL**	**LegalRuleML**	**ODRL**	**L4LOD**	**ELI**
Extended name	European Vocabulary	Creative Commons Rights Expression Language	Legal Rule Modeling Language	Open Digital Rights Language	Licenses for Linked Open Data	European Legislation Identifier
Legal domain	cross-domains	licences	cross-domains	policies	licences	cross-domains
Purpose	indexing of the documentary information of the EU institutions	machine-readable standard to express licensing terms	modelling of legal norms allowing legal reasoning	representation of the conditions of usage of digital assets	representation of existing licensing terms in the Web of Data	metadata for the description of legal documents issued by the EU and its Member States
Year	1984	2008	2011	2012	2013	2014
Current version	4.9	unique version	1.0	2.2	0.2	1.2
Licence	commercial or non-commercial use allowed providing appropriate acknowledgement	CC BY 3.0	OASIS Intellectual Property Rights Policy	W3C Community Contributor License Agreement (CLA)	CC-BY-SA	reuse allowed providing appropriate acknowledgment
Updates frequency	high (18 Dec 2020)	low (1 May 2008)	high (6 Apr 2020)	medium (15 Feb 2018)	low (10 May 2013)	medium (17 Jul 2020)
References	358	166	56	17	31	42
Link	http://bit.ly/2MY0TpM	http://bit.ly/2Lua4gp	http://bit.ly/2sxpskV	http://bit.ly/2J75JPj	http://bit.ly/2m40FSn	http://bit.ly/2NyUimC

Table 2.2: Classification of ontologies published from 1984 to 2014 according to the *general information* class of features. Last revision of the information contained in the table: April 2021.

	LOTED2	**PPROC**	**LDR**	**Data Protection Ontology**	**ELTS**	**NRV**
Extended name	not found	Public Procurement Ontology	Linked Data Rights	not applicable	European Legal TaxonomySyllabus	Normative Requirements Vocabulary
Legal domain	tenders and procurements	tenders and procurements	policies	privacy	consumer law	cross-domains
Purpose	indexing, search and retrieval of European public procurement notices	management of public procurements and the execution of contracts	representation of policies of Linked Data resources	Model the GDPR concepts, focusing on the obligation of the data controller in relation with the rights of the data subject	provide a conceptual structure to a domain terminology vocabulary	representation of normative requirements and rules as Linked Open Data
Year	2014	2014	2014	2015	2016	2017
Current version	unique version	1.0.0	unique version	unique version	not released	unique version
Licence	GNU GPL v3	CC BY-SA 4.0	CC BY 4.0	not found	not applicable	not found
Updates frequency	low (16 Jan 2014)	low (29 Oct 2014)	low (1 Sep 2014)	low (16 Feb 2016)	not applicable	low (last update not found)
References	39	31	4	36	22	20
Link	`http://bit.ly/2m5os4q`	`http://bit.ly/2MWxGPq`	`http://bit.ly/2KU59cx`	`http://bit.ly/2uhumDv`	not applicable	`http://bit.ly/2KFwkIC`

Table 2.3: Classification of ontologies published from 2014 to 2017 according to the *general information* class of features. Last revision of the information contained in the table: April 2021.

	GDPRov	**GDPRtEXT**	**PrOnto**	**GConsent**	**DPV**
Extended name	not found	GDPR text extensions	Privacy Ontology	not found	Data Privacy Vocabulary
Legal domain	privacy	privacy	privacy	privacy	privacy
Purpose	represent the provenance of the consent and data life-cycle	representation of the GDPR concepts with direct links to its text	representation of the GDPR concepts for legal reasoning and compliance checking	represent information related to the consent given for processing personal data	provision of a vocabulary of privacy terms for promoting interoperability
Year	2017	2018	2018	2019	2019
Current version	unique version	0.7	not released	0.5	0.2
Licence	CC by 4.0	CC by 4.0	not applicable	CC by 4.0	W3C Community Contributor License Agreement
Updates frequency	high (31 Mar 2020)	low (12 Feb 2019)	not applicable	low (25 Nov 2018)	high (13 Jan 2021)
References	10	46	61	13	13
Link	`http://bit.ly/3pBf8E9`	`http://bit.ly/3rJTeke`	not applicable	`http://bit.ly/3uDdmVC`	`http://bit.ly/3tae4JA`

Table 2.4: Classification of ontologies published from 2017 to 2019 according to the *general information* class of features. Last revision of the information contained in the table: April 2021.

2.3.2.2 The representational dimension

The eleven features contained in this class concern all the modelling choices which are immediately reflected in methodologies and standards used to build the ontologies.

The *language* feature refers to the main natural language used to specify the concepts, the relations and the lexicon inside the ontology. The *development* feature indicates the approach adopted in the ontology building process, i.e. a bottom-up approach (from lexicon to concepts), a top-down approach (from legal foundations to lexicon) or a middle-out approach, which merges the techniques of the previous two methods.

The *construction* feature specifies if the modelling of the concepts and the relations of an ontology was performed manually or using some Natural Language Processing (NLP) technique to partially automatise the process of building the ontology. Linked to this aspect, two features concern the sources from which the concepts inserted in the ontology were chosen. The first one is *knowledge source (KS) for terms extraction*, that is legal documents or websites used to extract the relevant concepts and the corresponding ontology lexicon. In contrast, the *external vocabulary (EV) reference* feature refers to the existing ontologies and vocabularies which the ontology reuses specifying the URIs of some of its concepts and properties. Therefore, the difference between these two last features is that the legal documents listed in correspondence of the first feature only provide the raw concepts which are relevant for the domain but which needed to be formally modelled before being inserted in the ontology. By contrast, the second feature looks at the reuse of some parts of existing ontologies in order to adopt some concepts and relations already modelled by them. Similarly, the *ground ontology* feature refers to the main ontology which is extended by the analysed resource. This feature can be seen as a specialisation of *external vocabulary reference*. The difference is that an ontology which uses another one as ground ontology inherits from it the great part of its concepts and structure, while an ontology that makes some references to external vocabularies adopts its own structure and reuses only some concepts of other existing resources.

The *level of structure* feature is a quantitative evaluation of the number of concepts and relations modelled by the ontology. This property can be expressed by three values that denote a growing number of classes and relations: *lightly structured*, *moderately structured* and *highly structured*. The *knowledge representation (KR) formalism* refers to the formal language used to represent the ontology in a machine readable way. At present, the two *de facto* standards used to represent ontologies are RDF and OWL. Connected to this feature, the *axioms* feature is also considered. It refers to the three possible level of axioms allowed by the OWL 2 specification: class expression axioms, object property axioms and data property axioms.

Taking into account the principle of reuse promoted by the Semantic Web, the *ontology design patterns* feature is used to represent some parts of knowledge whose modelling was already codified in a standard representation. Finally, the *evaluation* feature analyses which methods were adopted to evaluate the created knowledge model provided by the ontology.

Tables from 2.5 to 2.7 classify the analysed ontologies according to the features of this class.

	Eurovoc	**ccREL**	**LegalRuleML**	**ODRL**	**L4LOD**	**ELI**
Development	not found	not found	not found	not found	not found	not found
Construction	manual	manual	manual	manual	manual	manual
Language	EU's languages, Macedonian, Albanian, Serbian	English	English	English	English	English
KS for terms extraction	ECLAS thesaurus, SCAD, EC-01, Official Gazette indices	not found	not applicable	not applicable	not found	not applicable
EV references	FRBR, Dublin Core, SKOS	not found	not found	Dublin Core, SKOS, FOAF	not found	FRBR, Dublin Core, SKOS
Ground ontology	none	none	RuleML	none	none	FRBR/RDA
Level of structure	lightly structured	lightly structured	highly structured	highly structured	lightly structured	lightly structured
KR formalism	RDF	RDF	RelaxNG and XML Schema, RDFS, XSLT	RDF	RDF	OWL
Axioms	not found	class level, property level	class level, property level	class level, property level	class level	class level
Design patterns	not found	not found	container, collection, recursive element, marker, composite	not found	not found	not found
Evaluation	EU institutions, Publication Office, national and regional parliaments	not found	not found	not found	not found	provided by users

Table 2.5: Classification of ontologies published from 1984 to 2014 according to the *modelling information* class of features. Last revision of the information contained in the table: April 2021.

	LOTED2	PPROC	LDR	Data Protection Ontology	ELTS	NRV
Development	middle-out	bottom-up	not found	bottom-up	bottom-up	not found
Construction	manual	manual	manual	manual	manual	manual
Language	English	English	English	English	EU's languages	English
KS for terms extraction	TED website, EU Directive 2004/17/EC, EU Directive 2004/17/EC	buyer profiles, EU directives, public procurements' announcement models of Spanish legislation	not found	GDPR, Data Protection Directive, Handbook on European data protection law	vocabulary for the Uniform Terminology project	not applicable
EV references	LKIF-core, GoodRelations	CPV, PCO, FOAF, SKOS, DC, Organization Ontology, schema.org, Good Relations	ODRL, SKOS	LKIF-Core, SKOS	not found	LegalRuleML, RuleML
Ground ontology	none	none	ODRL	none	none	LegalRuleML
Level of structure	moderately structured	highly structured	lightly structured	lightly structured	lightly structured	moderately structured
KR formalism	OWL	OWL	OWL	OWL	not used	RDF
Axioms	class level	class level	class level, property level	class level	not applicable	class and property level
Design patterns	Social Reality	not found	not found	not found	not applicable	not found
Evaluation	not found	provided by two Spanish public authorities	not found	not found	legal experts in the loop	SPARQL queries

Table 2.6: Classification of ontologies published from 2014 to 2017 according to the *modelling information* class of features. Last revision of the information contained in the table: April 2021.

	GDPRov	**GDPRtEXT**	**PrOnto**	**GConsent**	**DPV**
Development	bottom-up	bottom-up	top-down following the MeLOn methodology	bottom-up following the "Ontology Development 101" methodology	bottom-up following the NeOn methodology
Construction	manual	manual	manual	manual	manual
Language	English	English	English	English	English
KS for terms extraction	GDPR	GDPR, document issued by official sources, industry-based sources	GDPR, terms of use, information, privacy policies, consent forms	institutional and, academic resources	GDPR, EnterPrivacy
EV references	not found	ELI ontology	ALLOT, FRBR, LKIF Core, PWO, LegalRuleML metamodel	GDPRov, GDPRtEXT Time Ontology	Dublin Core
Ground ontology	Prov-o, P-Plan	ELI ontology	none	none	Special Usage Policy Language
Level of structure	lightly structured	lightly structured	highly structured	highly structured	lightly structured
KR formalism	OWL	OWL	OWL	OWL	OWL
Axioms	class level	class level	class level, property level	class level, property level	class level
Design patterns	not found	not found	Time-indexed Value in Context, Time interval	not found	not found
Evaluation	SPARQL queries	not found	SPARQL queries	SPARQL queries	feedbacks expected from users

Table 2.7: Classification of ontologies published from 2017 to 2019 according to the *modelling information* class of features. Last revision of the information contained in the table: April 2021.

2.3.2.3 The semantic dimension

The two sets of features presented so far are independent from the legal domain and they could be applied potentially to analyse and compare the ontologies belonging to every domain of interest. By contrast, the three features belonging to this class specifically refer to the way in which the legal knowledge is modelled.

The *modelling of temporal aspects* feature specifies if an ontology models some temporal aspects concerning the legal field of interest and provides a brief description of the way in which this is done. There are a lot of different possibilities to model a temporal feature inside an ontology: it could be a simple time mark associated to the issue of a policy, or an interval of time which specifies the validity of an obligation or it could be an implicit representation of time which focuses on the parameters that could vary over it, e.g., the status of a norm or the jurisdiction under which it is valid.

When an ontology models norms and rules, the *adopted normative model* feature specifies the type of rules that the ontology can represent, i.e. constitutive and prescriptive (or regulative) norms, as defined in [24]. Finally, the *deontic logic model* feature provides a short description of the deontic operators modelled inside the ontology, i.e.obligation, duties, permissions and rights. As for the previous feature, this one holds only if the ontology deals with norms and rules. However, since norms are one of the main focus of the legal domain, a lot of the analysed ontologies model the deontic operators. For example, some of them only represent permissions, obligations and prohibitions, others model also the violations of obligations and prohibitions, while others provide a hierarchy of deontic operators organising them according to different criteria (e.g., temporal criteria or need for compensation of a violated norm).

Tables from 2.8 to 2.10 classify the ontologies according to these three features.

	Eurovoc	**ccREL**	**LegalRuleML**	**ODRL**	**L4LOD**	**ELI**
Modelling of temporal aspects	not applicable	not found	modelling of the aspects of a rule that vary over time (e.g., status, validity, jurisdiction)	modelling of the date and time a policy is issued or modified. Date and time constraint on the validity of a deontic operator.	not found	not applicable
Adopted normative model	not applicable	prescriptive rules	constitutive and prescriptive rules	prescriptive rules	prescriptive rules	not applicable
Deontic logic model	not applicable	requirements and prohibitions set by the Creative Commons standard	permission, rights, obligation, prohibition, compliance with a prohibition or an obligation, violation of a prohibition or an obligation, reparation of a violation	permissions, prohibitions and obligations over a digital or material asset	permissions, obligations and prohibitions over the licensed data	not applicable

Table 2.8: Classification of ontologies published from 1984 to 2014 according to the *semantic information* class of features. Last revision of the information contained in the table: April 2021.

	LOTED2	PPROC	LDR	Data Protection Ontology	ELTS	NRV
Modelling of temporal aspects	date and time associated to tenders	only to indicate the deadline for submissions of tenders and requests of participation	not found	not found	use of the ontological relation “replaced by” to model the evolution of concepts definitions	temporal aspects are modelled through the concepts of perdurance, persistence, co-occurrence and preemptiveness of a deontic operator
Adopted normative model	not applicable	not applicable	prescriptive rules	prescriptive rules	constitutive and prescriptive rules	prescriptive rules
Deontic logic model	not applicable	additional obligations that a contract needs, requirements that a tender needs in order to be submitted	right over a Linked Data resource	obligation (of the data controller) and rights (of the data subject)	permissions, prohibitions, obligations and rights are complex concepts that are linked to other concepts they involve (e.g. “right to withdrawal” is linked to the concept “withdrawal” intended as an act)	permissions, obligations and prohibitions are organised according to the principles of compensation, compliance, violation, temporal validity and realisation

Table 2.9: Classification of ontologies published from 2014 to 2017 according to the *semantic information* class of features. Last revision of the information contained in the table: April 2021.

	GDPRov	GDPRtEXT	PrOnto	GConsent	DPV
Modelling of temporal aspects	the temporal dimension is implicit in the modelling of processes	information about personal data retention and storage period are modelled as ontology classes	temporal intervals associated to actions in workflows, to agents' roles and to deontic operators	temporal values for representing the state of consent	not applicable
Adopted norms model	prescriptive rules	prescriptive rules	constitutive and prescriptive rules	not applicable	not applicable
Deontic logic model	not found	obligations of different agents mentioned in the GDPR (Controller, Processor, Data Protection Officer) and right of the data subject	permissions, prohibitions, obligations, rights, compliance with an obligation and violation of an obligation. Some references are modelled between obligations and compliance/violation and between rights and permissions	not applicable	not applicable

Table 2.10: Classification of ontologies published from 2017 to 2019 according to the *semantic information* class of features. Last revision of the information contained in the table: April 2021.

2.3.3 Concluding remarks and open challenges in the representation of legal knowledge

From the analysis of the ontologies contained in Section 2.3.2 and for each macro-class of feature used to classify them, it is possible to identify some issues and future challenges to be addressed in the field of legal knowledge representation.

Considering the general information about an ontology (summarised in the *informational dimension* of the proposed analysis), some lacks of standardisation still exists in the graphical user interfaces (GUIs) used to make the ontology content available to the final user. Currently, the LODE[47] tool is one of the most common Web services used to automatically create these GUIs. LODE processes the OWL file of an ontology to create an HTML page which lists classes, properties and axioms of the ontology together with some metadata indicating the author(s), the release date, the current version and the licence of the ontology. A unified look for the GUIs exposing the content of an ontology could be helpful for users concerned with ontology building and reuse, as it could reduce the time spent to look for the information within websites creating over time a kind of "familiarity" with the interface, by knowing exactly the way in which the information is organised.

Linked to this problem, the second issue is related to the need of making explicit all the details concerning the download and the licence of an ontology. Browsing the Web pages of the different resources, it is sometimes difficult to find this information. However, it seems clear that without them, a fair reuse of the ontologies would not be promoted. A special case concerns the resources made available by the European Union whose orientation towards the Semantic Web and the Linked Open Data is remarkable. They are all collected in the EU vocabularies portal[48] where a tab-like GUI organises all the information about a resource. However, even if the download links are well visible, the type of licence which regulates the use of each resource is not specified. Moreover, in the current interface of the EU vocabularies portal, the title of each tab sometimes does not clarify the information associated with it, and the documentation of the different resources is not standardised. For example, the documentation of ELI is a PDF file which contains few information about the ontology. In contrast, the description of Eurovoc is better organised into expandable windows inside the tab. Therefore, according to these remarks, some improvement would be desirable to harmonise the way in which the metadata on legal ontologies issued by the EU are organised inside the portal.

[47] github.com/essepuntato/LODE
[48] publications.europa.eu/en/web/eu-vocabularies

Concerning the methodological and technological choices made during the development of an ontology, this information is never displayed on the aforementioned GUIs and it could be difficult to find it also reading the literature published together with the ontology. However, this information is important for several reasons. First, it provides a scientific foundation to the work allowing other researchers to analyse and verify it. Second, it enables an easy and understandable interpretation of the corresponding literature in which this information is sometimes implicit, even if it is at the basis of the development of the ontology.

The analysed resources show a positive trend towards the reuse and extension of concepts and properties modelled in other existing ontologies, while there is still a lack of sensitivity to the adoption of the ontology design patterns (see Section 2.2.3) in the ontology building process. The low use of ontology patterns could be associated with the difficulty to identify, inside a complex modelling problem, the parts which could be covered by an ODP because that requires the knowledge of the full landscape of available ODPs.

Finally, the classification of the ontologies according to the *representational dimension* reveals a lack of standard methodologies to evaluate the proposed knowledge models (as it is evident also from Section 2.1.4). In the reference academic papers of some resources, the criteria used to evaluate the proposed models are sometimes omitted. However, as it can be noticed in Table 2.6 and 2.7, the current trend is to provide SPARQL queries to test the validity of some competency questions and the fulfilment of some objectives which an ontology should reach. This approach is especially adopted by the most recent ontologies as, for instance, NRV and PrOnto. By contrast, older ontologies mention in their literature the fact that they are adopted by real users, as in the case of PPROC or the resources released by the European Union.

The weaknesses concerning the *semantic dimension* call back the aforementioned problem of the ontologies design patterns. Indeed, each ontology models a specific legal domain and adopts its own ontological commitment, with a consequent proliferation of different knowledge models referring to similar use cases. For instance, the deontic operators, being one of the main focus of different legal domains, are represented in many ontologies, but the aspects that each of them considers are different. Some ontologies associate a temporal reference to the validity of an operator (as LegalRuleML or ODRL do) while others do not (e.g. L4LOD). Furthermore, some ontologies make a distinction between an obligation which is respected and an obligation which is violated (as NRV), while others not (e.g. LDR). Consequently, even if recurrent use cases could be possibly identified within the legal domain, few efforts are dedicated to find standardised and extensible solutions to them.

According to the proposed three-dimensional analysis and the identified weaknesses in existing legal ontologies, some improvements could be done to enhance the ontology building process towards the reuse of existing resources.

First of all, the identification of a recommended set of metadata to include inside the ontology source file should be evaluated in order to complete the information that is already shown in the graphical interfaces displaying the content of an ontology. Some metadata for representing the methodology utilised for developing the ontology and, eventually, the adoption of existing design patterns would be useful to ensure the reuse of the ontology itself. Moreover, a set of metadata able to summarise some of the purely legal aspects modelled into an ontology could be envisioned. Some of these metadata could recall the features used inside this chapter to classify the ontologies, like the modelled deontic operators and the type of modelled norms (if this feature is applicable).

In addition to a recommended set of metadata for the description of the ontology features, it could be important to address the need for legal design patterns. An effort to discover recurrent legal knowledge and to model it in the form of standardised legal use-cases could improve the quality of the released ontologies, reducing the efforts needed to model legal knowledge. This is especially true considering that the design of ontology-based systems is usually assigned to heterogeneous teams, which include both legal experts and computer scientists. When starting the development of a new ontology, the existence of ontological patterns for modelling legal knowledge could enhance the interdisciplinary dialogue between legal and technical experts, providing a common ground for discussion.

2.4 An applicative experience for a controlled exploration of ontologies: InvestigatiOnt

This section describes a Web application that has been developed based on the analysis of the legal ontologies proposed in Section 2.3. Despite this is not the main focus of the thesis, the aim of this application is to show how a theoretical analysis could trigger the development of new tools for discovering the features of existing legal ontologies. Specifically, considering the diverse landscape of ontological resources released in the last decade for modelling legal knowledge, the objective of this Web application is to support the end-users in a guided exploration of those resources, for understanding their ontological commitments and, eventually, promote their reuse.

2.4.1 Motivations

As mentioned in Section 2.3.3, the design of legal ontologies is usually assigned to heterogeneous teams of experts from both the legal and the technological fields. On the one hand, the legal domain is characterised by some complexities that could be difficult to be addressed by technological experts who lack a legal background. Some of those complexities could concern: *(i)* the existence of different legal systems (e.g., common law, civil law) and jurisdictions (e.g., local, national, international), *(ii)* the way the norms could interact (e.g., norms that express the obligation to be accomplished if another obligation is violated, or norms that express an exception to an obligation) or *(iii)* the structure that characterises the legal texts (e.g., their division in articles, paragraphs, definitions). On the other hand, legal experts who lack a background about semantic technologies may experience some difficulties in understanding the modelling choices of a legal ontology which formalised in a machine-readable representation language.

In this context, there is the need to define tools able to support both technological and legal experts towards a better understanding of the legal concepts expressed in the ontologies. Based on this consideration, the developed Web application, called *InvestigatiOnt*, aims to support the interested user to explore the ontologies that have already been modelled in different legal domains and possibly to choose the one that better suits her modelling requirements. To achieve this objective, InvestigatiOnt offers two types of service: the *visualisation service* displays the information concerning an ontology, and the *search service* suggests one or more ontologies suitable to meet the user's requirements analysing the answers she provided to a set of questions.

In the implemented demo, InvestigatiOnt supports to exploration of 12 ontologies belonging to six different domains (slightly different from the grouping proposed in Section 2.3):

1. *legal norms*: the ontologies (LegalRuleML and NRL) model the norms as they could be found in the legal documents issued by local, national or international governments;

2. *policies*: the ontologies (ODRL and LDR) model the permitted, mandatory and prohibited actions that can be made on a digital or material asset;

3. *licences*: the ontologies (CCRel and L4LOD) model the actions allowed on a resource protected by the intellectual property right;

4. *legal documents representation/indexing*: the ontologies (Eurovoc and ELI) represent the text structure of legal documents and their topics;

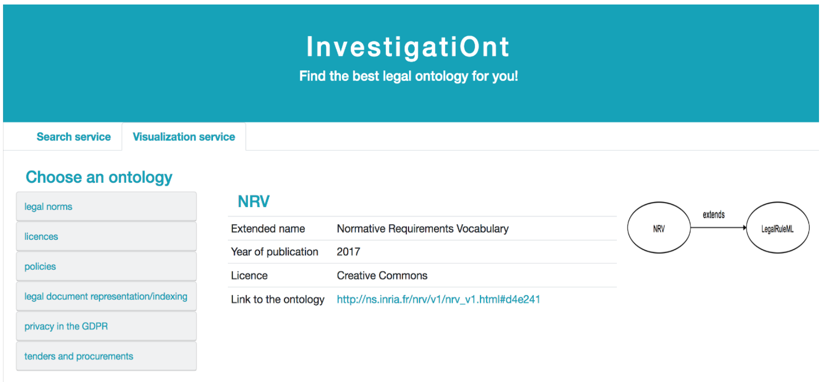


Figure 2.3: A screenshot of the interface for the visualisation service.

5. *privacy in the GDPR*: the ontologies (GDPR and the Data Protection Ontology) model the concepts involved in the new European General Data Protection Regulation;

6. *tenders and public procurements*: the ontologies (LOTED and PPROC) model the processes used by the public administration to find contractors to entrust with services or supplies.

The following section describes the services implemented in InvestigatiOnt for exploring the aforementioned ontologies.

2.4.2 The InvestigatiOnt services

2.4.2.1 The visualisation service

The visualisation service displays the basic information about an ontology. More precisely, for a selected ontology, the user visualises the following information: *(i)* the *extended name of the ontology* and its acronym; *(ii)* the *year of publication* of the ontology; *(iii)* the *last update* of the resource (if this information is not displayed, the publication year coincides with the last update); *(iv)* the *licence* under which the ontology is made available for re-use, and *(v)* the *link* to the official documentation of the ontology. In addition to these types of information, also a chart showing the dependencies among different ontologies is displayed. In particular, the char shows two types of relation: *extends*, indicating that the selected ontology further specialises some of the concepts in the other ontologies, and *re-uses*, indicating that the selected ontology reuses concepts modelled in the other ontologies, without however specialising them. A screenshot of the implemented interface for the visualisation service is shown in Figure 2.3.

The information about an ontology provided in the visualisation service recalls some of the features identified in the feature-based analysis proposed in Section 2.3, specifically in the informational and in the representational dimensions. As a future work, the information provided by the visualisation service could span all the features identified for an ontology in the two aforementioned dimensions. By contrast, the semantic dimension of an ontology could be explored through the search service, as it will be explained in the following subsection.

2.4.2.2 The search service

The search service supports the user in the exploration of existing ontologies by suggesting her the one that better fits her requirements. To do this, the user is asked to answer a list of questions. Each question has a closed set of possible answers, and a response is required before moving to the next one. Each question is coupled with clarifying examples and aims to understand if and how a user needs to model a specific legal aspect in her domain of interest. Thus, through those questions, InvestigatiOnt tries to understand the ontological commitment the user wants to assume.

As shown in Figure 2.4, the first question asked to the user concerns the legal field that she needs to model, choosing among the six legal fields listed above. Depending on the selected answer, the next questions will vary, in order to understand which ontology of the selected legal field is more suitable to fulfil the user's requirements. The questions are formulated according to two different templates: *(i)* the question recalls a feature belonging to one of the ontologies of the chosen field and the user is asked whether this feature is necessary for her modelling requirements (possible answers: *yes* or *no*), and *(ii)* the question asks the user to choose the way she wants the legal field to be modelled inside the ontology she is looking for (this kind of answer is more complex and it requires examples to ease the choice). As previously mention, the research service has been conceived to explore the semantic dimension of the ontologies, as it was discussed in the feature-based analysis proposed in Section 2.3. Consequently, the questions have been drafted to investigate the three features proposed for that dimension.

As the user goes on by answering the questions, the interface of InvestigatiOnt changes displaying further information. The first one is a track of the previous answers provided by the user. This is intended to help her to remember the selected answers, allowing her (if needed) to go back and change the answer to one or more questions. The second one is the information about the available ontologies in the legal field chosen at the first step. In particular,

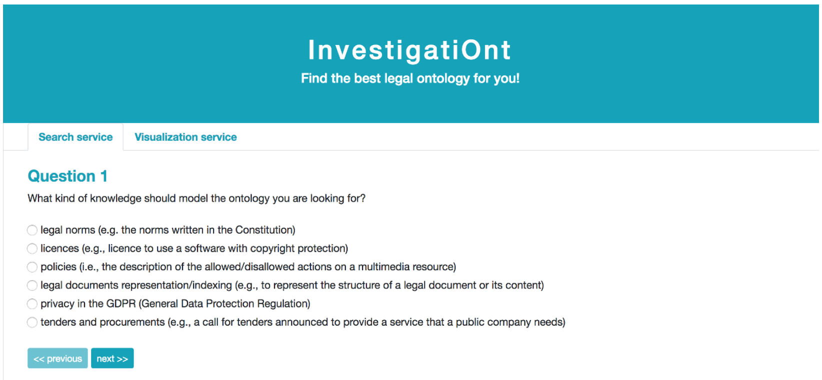


Figure 2.4: The first question of the search service.

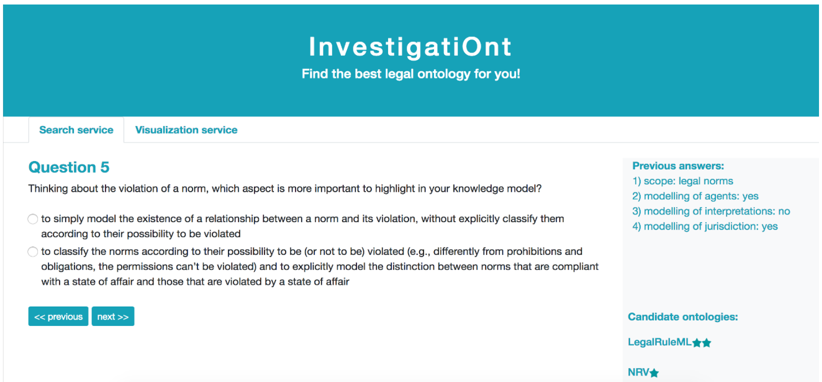


Figure 2.5: The interface of the search service during the interaction.

every time a user answers a question, a score is assigned to one of the available ontologies. The basic idea is that each question tries to discriminate on a specific feature of the ontologies belonging to the selected legal field. Thus, at each step, a unitary score (graphically displayed as a star) is assigned to one of the ontologies and it is added to the scores accumulated through the previous steps. A screenshot of the interface displaying this information while answering the questions is shown in Figure 2.5. Through this mechanism, the user is supported in her selection process with a fully transparent process that leads to the final recommendation for an ontology, provided by the system.

As a result, the system shows a summary table, similar to that shown in Figure 2.6. In this table, each row corresponds to a feature for which the user provided a response, while each column represents an ontology. A check symbol in a cell indicates that an ontology models the corresponding feature. Consequently, this table provides an explainable recommendation to the user,

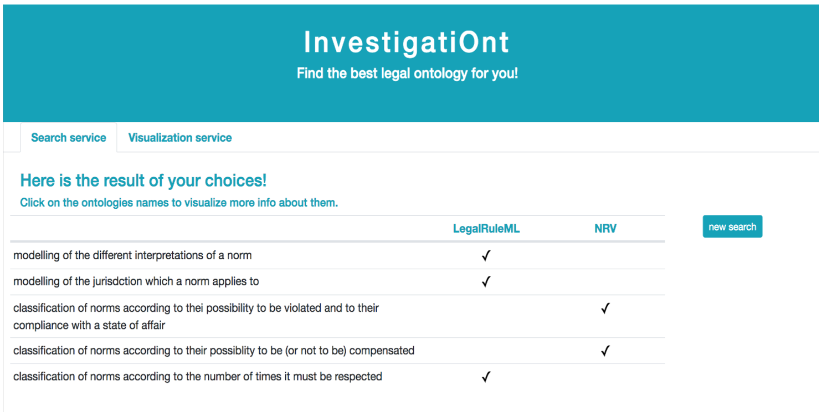


Figure 2.6: The interface of the search service when all the answers to the were provided and the results are displayed.

that could be useful especially when the scores between two or more ontologies are similar because the user has the possibility to evaluate the pros and cons of the choice of one resource rather than the other. Moreover, to further help her in the evaluation, clicking on the name of an ontology, the user is redirected to the visualisation service, which provides more technical information about the ontology.

2.5 Summary

This chapter presented the technological background of the thesis. The first part of the chapter provided an overview of the different interpretations of the term *ontology* and presented the main aspects involved in the ontology engineering process. The Semantic Web was presented with a focus on its effort to promote the interoperability and sharing of knowledge, through the provision of standards (RDF, RDFS, OWL, SKOS) for the representation of knowledge.

The second part of the Chapter focused on the first contribution of this thesis, i.e. an analysis of existing legal ontologies developed at three levels of comparison. This analysis highlighted the variety of possible representations of legal knowledge. The applicative outcome of this analysis was the development of a Web application that helps users interested in the reuse of existing legal ontologies to explore the variety of those knowledge models.

3 The Protection of Personal Data at the European Level

This chapter presents the legal context of this thesis, i.e. the data protection domain with a focus on the GDPR.

The first part of the chapter provides a historical overview of the development of the right to data protection in the European Union. Then, it describes core concepts of the legal framework set by the GDPR, with a particular focus on Articles 12-14. Those articles concern the principle of transparency and the obligations for the data controller to provide individuals with the information about the processing activities performed on their personal data. With a reference to this principle, the second part of the chapter discusses some of the characteristics of privacy policies that, as the main means of communication used by the data controller to comply with the principle of transparency, undermine the right of the data subject to receive information about the processing activities performed on her data.

3.1 Data Protection Law in the European Union

3.1.1 An Overview of the Historical Development of the Right to Data Protection

In the European legal framework, the right to the protection of personal data has come to shape as a result of the development of the modern society. In 1950, this right was encompassed within a broader right to privacy, i.e. the right to respect for private and family life, home and correspondence, as described by Article 8 of the European Convention on Human Rights[1] (ECHR), entered into force in 1958.

Later on, with the development of the information society, a new concept

[1] Council of Europe, *European Convention on Human Rights*, CETS No. 005, 1950.

of "informational autonomy" (or "self-determination") emerged to express the individuals' right to decide which of their personal information may be disclosed, to whom and for which purpose [51]. In light of this new perspective on data concerning individuals, in 1981 the Council of Europe Convention 108[2] (Convention 108) acknowledged in Article 1 the right of individuals to the protection with respect to the processing of their personal data. Convention 108 is, to date, the only legally binding international instrument in the field of data protection and it lays at the core of any national legal framework in this field. Its main principles, outlined in Article 5, focus on the guarantee of a fair and lawful collection and processing of personal data, limiting the possibility to store personal data to those that are adequate and relevant for a specified and legitimate purpose and just for the amount of time that is necessary for accomplishing the purpose of the storage. The adoption of appropriate security measures for stored personal data is advocated in Article 7, while Article 8 lists additional safeguards for individuals, as the right to access, rectify or erase their personal data. In 2018, the Convention 108, initially drafted as a technologically-neutral data protection instrument, underwent a process of modernisation, with the aim of adapting to the new reality of the digital world and the emergence of new data practices [43]. The modernised version of the Convention 108[3] introduces new rights for the individuals and increases the responsibilities for the entities that process personal data.

At the time of its entry into force, the Convention 108 played a fundamental role in defining the ultimate distinction between the right to privacy and the right to data protection. While, on the one hand, the right to privacy concerns the situations of interference with private life and the compromise of certain information that could impact public opinion against an individual, on the other hand, the right to data protection is a "modern and active" right that protects individuals whenever their personal data are processed, regardless of the impact on their privacy. [44].

In the EU primary law, the right to data protection entered as a "third generation" right to reflect the modern society [67] in Article 8 of the Charter of Fundamental Rights of the European Union[4] (the Charter). The Charter became a legally binding document in 2009 with the entry into force of the Lisbon Treaty[5], that recognises the protection of personal data as a fundamental

[2]Council of Europe, *Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data, CETS No. 108, 1981.*

[3]Council of Europe, *Protocol amending the Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data*, CETS No. 223, 2018.

[4]European Union (2012), *Charter of Fundamental Rights of the European Union*, O.J. 2012 C 326.

[5]European Union (212), *Consolidated version of the Treaty on the Functioning of the European*

right and provides specific legal basis for the European Union to act in the data protection field under Article 16.

At the level of EU secondary law, the Data Protection Directive[6] entered into force in 1995 as an effort made by the European Commission to harmonise the Member States' data protection legal frameworks that were rather fragmented and inconsistent, despite the guiding principles set out in the Convention 108[7]. The harmonisation effort of the Directive is witnessed by the introduction of independent supervisory authorities responsible for monitoring compliance with national law on the territory of competence, handling complaints and providing consultation. The cooperation of the representatives for these national authorities resulted in the EU "Article 29 Working Party" advisory board.

In 2009, launching a public consultation, the European Commission engaged in a process of update of the legal framework on data protection [68]. The motivations that led to the beginning of this reform were many. On the one hand, there was a desire to keep track of the latest technological advancements, dealing with the new value of data as a driving factor in businesses development and facing the new privacy risks arising from an increasingly automated data handling. On the other hand, there was a need for further efforts of harmonisation, which had faded away in the transposition of the Data Protection Directive into the various Member States laws [29]. The reform of EU data protection legislation was achieved with the adoption of the GDPR[8] in April 2016 and its entry into force on 25 May 2018. The decision to replace the Data Protection Directive with a regulation, that is immediately applicable in all Member States without requiring additional implementation efforts, makes clear the harmonisation objective pursued by the European Commission in aligning the data protection legal framework among Member States. While maintaining the core principles of the Data Protection Directive, the GDPR boosts the individuals' rights to the protection of their personal data and provides new obligations to organisations that process personal data, potentially increasing fines in case of non-compliance. The core concepts set by the GDPR are discussed in the

Union, O.J. 2012 C 326.

[6]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, O.J. 1995 L 281.

[7]European Commission (1990), *Commission Communication on the protection of individuals in relation to the processing of personal data in the Community and information security*, COM(90) 314 final 90/C 277/05.

[8]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), O.J. 2016 L 119.

next Section.

3.1.2 The General Data Protection Regulation

The GDPR provides a broad definition of its territorial scope. According to Art. 3, it applies to all businesses that are established in the EU Member States. Moreover, regardless their place of establishment, it applies to businesses that offer their good/services to data subject in the EU or monitor their behaviour, when it takes place within the Union. Under those conditions, the GDPR sets the legal framework for the processing of personal data. The main aspects of the Regulation that are relevant for this research work are highlighted in the following sections.

3.1.2.1 Definition of Personal Data

The Regulation defines a *personal data* as any information that could identify, directly or indirectly, a living individual (the *data subject*) (Article 4(1)). Consequently, this definition includes all unique identifiers, utilised in different situations of everyday life, that are commonly associated with a person, such as the name, the telephone number, some location data or the identifiers used in online contexts. However, the definition is not limited to those identifiers. By contrast, a data is *personal* when the fragment of information that it represents, combined with other fragments of information, could unravel the identity of the data subject, making her distinguishable from other individuals.

Among the personal data, some of them can be processed only in a limited set of circumstances. This is the case for the *special categories* of personal data, i.e. those that could reveal the racial or ethnic origin of the data subject, as well as her political opinions, beliefs, trade union membership, health and sexual life. Generic and biometric data are also considered as special categories of personal data, when they are used for the identification of the data subject.

Any operation performed on those personal data is referred to as *processing*. Article 4(3) mentions processing actions such as: collection, recording, storage, consultation, use, disclosure by transmission, alignment, combination or erasure. The GDPR is technology neutral and protects personal data regardless of the technology used for the processing. Consequently the GDPR applies also to the personal data that are processed manually, as long as these data are contained in a filing system organised and accessible according to some systematic criteria (e.g. the alphabetical order).

3.1.2.2 Actors involved in the Data Processing

The processing of personal data involves actors other than the data subject. Specifically, the *controller* is the entity responsible for determining the purposes and the means of the processing, i.e. “why” and “how” personal data are processed (Article 4(7)). When the purposes and the means of the processing are jointly determined by more than one entity, these entities are called *joint controllers* (Article 26). Furthermore, the entity that processes the personal data on behalf of the controlled is defined as *processor* by Article 4(8). Thus, the role of processor is determined by the controller, who decides whether to process personal data within its organisation or appointing an external organisation as a processor, delegating to it some of all part of the processing [192]. All the entities the personal data are disclosed with are called *recipients* (Article 4(9)). Among the recipients, *third parties* are defined by Article 4(10) as those entities other than the data subject, the controller, the processor and the persons acting under the direct authority of the controller of the processor. In some circumstances listed by Article 37, the data controller has the duty to appoint a *data protection officer (DPO)* with advisory and monitoring tasks related to the compliance with the GDPR. The DPO should also act as an intermediary with the national *data protection authority (DPA)*, i.e. an independent public authority that supervises the application of the data protection law and handles complaints lodged against violations of the General Data Protection Regulation and the relevant national law (Articles 51 to 59) [66].

3.1.2.3 Principles of the Data Processing and Lawful Grounds

The processing of personal data should follow some principles set in Art. 5. Specifically, the processing of personal data should be undertaken *lawfully* with respect to the legal grounds set in Art. 6(1). In order for the processing of personal data to be lawful, it must comply with one of the following *legal bases* (also called *legal grounds*): the data subject provided her consent for the processing; the processing is necessary to enter into or perform a contract with the data subject; the processing is necessary to protect the vital interest of some individuals or to comply with a legal obligation; the processing is necessary for public interest under EU or national law; the processing is necessary for the legitimate interest of the controller or the third party, when the processing doesn’t impact on the fundamental rights and freedoms of the data subject.

The principle of *fairness* asks the data controllers to be able to demonstrate the compliance with the Regulation when performing processing operations on personal data. Moreover, data controllers should notify data subjects about the processing of their personal data, in a way that makes them aware of potential risks deriving from the processing activity [45].

The principle of *transparency* asks the controller to provide the data subject with information about the processing activities performed on her data. This information is usually described in privacy policies. Art. 13 and Art.14 provide a detailed list of information do be provided, as will be discussed in Section 3.1.2.5

The principle of *data minimisation* asks for processing activities to be performed only on the personal data that are necessary for accomplishing the purpose of the processing, ensuring the *accuracy* and, when necessary, the update of the data being processed. Moreover, the processing of personal data should ensure appropriate security measures to avoid unauthorised processing or accidental loss *(integrity and confidentiality)*.

The principle of *purpose limitation* requires data to be processed without diverging from the original purpose for which they were collected, while the *storage limitation* sets the storage of personal data for the limited amount of time that is necessary to fulfil the purpose of the processing.

Finally, according to the *accountability* principle, the controller should be able to demonstrate the adherence to the aforementioned principles.

3.1.2.4 The Rights of the Data Subject

As anticipated in Section 3.1.1, the GDPR boosts the rights of the data subjects. Specifically, data subjects have the right to be informed about the processing of their personal data, obtaining the access to the processed data. In case of inaccurate or incomplete information, the data subject has the right to request a correction (i.e. rectification) of this information. He can also asks for the erasure of the data when they are no longer necessary for accomplishing the purpose of the processing or the processing is unlawful. In specific cases, the data subject has the right to object to the processing of his personal data or to restrict the processing. He equally has the right to requests the intervention of natural persons in case of decision based on the automated processing, when the automated decision could significantly affect him. Lastly, the GDPR introduces a new right for the data subjects, providing them with the possibility to ask for their personal data in a machine-readable format and transmitting them to another controller. These rights apply across the EU, regardless of where the data is processed and where the company is established (also including non-EU companies that offer goods and services in the EU).

3.1.2.5 A focus on Articles 12 to 14: requirements for transparency

The principle of transparency and fairness, enshrined in Article 5(1)(a), are fundamental in regulating how the data controller communicates information to the data subject. Recital 60 of the Regulation states that those principles “require that the data subject to be informed of the existence of the processing operation and its purposes”. Moreover, according to recital 39, “natural persons should be made aware of risks, rules, safeguards and rights in relation to the processing of personal data and how to exercise their rights in relation to such processing”.

To ensure compliance with those principles, Article 12 obliges the controller to proactively provide the data subject with any information that she can easily read and understand, using “clear and plain language”, i.e. avoiding complex legal constructions and ambiguities in the statements [200]. Article 12 also points out that this information should be provided in an easily accessible form (like paper or electronic documents, some audio-visual format or orally), depending on the user needs [191]. The need for transparency of communications, thus, aims to address the hurdles that commonly affect the privacy policies (as it will be discussed in Section 3.2) and guarantees that the data subject receives complete and intelligible information enabling her to exercise her rights.

The data subject has the right to be informed about the processing of her data regardless the source the personal data come from, being it the data subject or not. Specifically, when the personal data are collected directly from the data subject, according to Article 13, the controller must specify the following information: (i) the contact details of the controller and of the DPO (if there is one), (ii) the purpose of the processing and its legal ground, (iii) the controller’s legitimate interest, when it is declared as the lawful ground for processing, (iv) the entities that eventually receive the collected personal data, (v) the intention of the controller to transfer the data outside the European Union, (vi) how long the personal data will be stored, (vii) the rights that the data subject can exercise with respect to the protection of her personal data, (viii) the adoption of automated decision making systems, providing information about the logic involved and the consequences of such a processing. The obligation to inform the data subject about the adoption of automated decision-making systems has the clear goal of tackling the rapid development of machine learning and intelligent systems that take advantage of the increasing availability of data on the Web to drive the development strategies of the organisation. As anticipated in Section 3.1.1, this was one of the motivations that led the European Commission to the reform of the data protection legal framework, because, as stated in recital 58 of the Regulation, the principle of transparency "is of particular relevance in situations where the proliferation of actors and the technological complexity of practice make it difficult for the data subject to know and understand whether, by whom and for what purpose personal data relating to him or her are being collected, such as in case of online advertising". When personal data have not been obtained from the data subject, Article 14 requires the controller to provide, in addition to the information listed by Article 13, the categories of personal data concerned and the sources of these data. Furthermore, in both Article 13 and 14, the principles of transparency and purpose limitation, oblige the controller to inform the data subject when the purpose of the processing changes from the one that originally motivated the collection of the personal data.

What emerges from Articles 12 to 14 is therefore a principle of transparency that must guide and shape every phase of data processing. Transparency is required: for the information that must be provided *before* the start of processing, for the information that must be available *during* the processing of personal data and for the information that has to be provided *upon request*, when the data subject exercises her right of access to the data concerning her [45].

3.2 Providing information on the Data Processing

A privacy policy is a public written document released by an organisation for giving some information about the data processing activities of the data controller. As discussed in the previous section, this information should be communicated in a *"concise, transparent [...] form, using clear and plain language"* (Art. 12). However, privacy policies are notoriously long, domain-specific documents that the user will hardly be able to read and understand. Many of the hurdles to a transparent communication of data processing activities stem from the written form of privacy policies, thus they may also be taken into account when applying techniques of NLP on their texts, since they could negatively affect the performance of those techniques. The next sections will address these issues, highlighting to what extent they impact on the core principle of transparency.

3.2.1 Length of the documents.

In 2008, the average length of the 75 most visited Web sites in America over a year was estimated in 1514 words, taking 244 hours of reading time, equivalent to 40 minutes a day [125]. Ten years later, another study with a similar objective estimated the average length of the 20 most used apps worldwide is equal to 3964 words, with a 58% increase with respect to the study proposed in 2008 [174]. In 2019, the Commissioner for Justice, Consumers and Gender Equality, Věra Jourová, stated that among the 60% Europeans who read privacy statements, only 13% of them read them fully, pinpointing the problem in the length and difficulty of understanding privacy policies [69].

3.2.2 Required educational level for readability.

The legal nature of the privacy policies has raised concerns about their readability, that depends on the expected ability of the targeted audience to understand the privacy policies' text [103].

The Flesch Reading Ease Score (FRES) [75] is one of the most popular statistical measures that has traditionally been used to provide a quantitative assessment of the readability of privacy policies, based on word and phrase length [7, 103, 186]. In contrast to those readability formulae, which can be easily computed by text processing techniques, other approaches assessed privacy policies readability through the close test [184], requiring human participants to fill blank spaces in privacy policies text fragments [72, 179]. Regardless of the approach applied for determining privacy policies readability, all the studies agree in revealing that privacy policies are often difficult to understand and require a high level of reading skills [109]. The aforementioned results have been recently confirmed in [70], pointing out that reading a privacy policy requires a educational level equivalent to high school or college. In 2019, "The New York Times" published a study of 150 popular privacy policies and it claimed that a college educational level could even be insufficient for reading many of the analysed documents [117].

The outcome of these studies is particularly relevant for companies, as easy-to-read privacy policies lead to a positive increase in individuals' trust in companies and a greater willingness to disclose their personal data [64]. Consequently, privacy policies should be informative and accessible to readers, avoiding the abuse of domain-specific legal terms, that may increase user overwhelm and discourage them from reading [130].

3.2.3 Intentional ambiguity.

Natural language is inherently ambiguous and this ambiguity could occur at different degrees, such as at the level of one polysemous word carrying several meanings, or at the level of a whole sentence for which multiple interpretations are possible [190]. When it comes to the legal domain, the ambiguity of natural language is even more evident and it is used intentionally in written norms and rules. This choice is meant to deal with the unpredictable circumstances that could rise from the variety of human activities and to allow further refinements of those norms and rules accordingly [63, 115].

Following similar motivations, in written privacy policies, companies describe the implemented data-handling practices using deliberately ambiguous terms so as to accommodate future changes in their data practices, without however needing to change their privacy policies [19]. Ambiguity appears in various forms, including *semantic ambiguity* (more than one interpretation for a statement in its provided context), *vagueness* (relative interpretations or borderline arising from a statement) and *incompleteness* (lack of sufficient details for determining the meaning of a sentence) [123].

In the writing of privacy policies, the different forms of ambiguity derive from specific terminological choices which include the use of conditionals (e.g., “as needed”, “as applicable”), generalisations (e.g., “mostly”, “’typically’), modality (e.g. “may”, “could”) and numeric quantifiers (e.g., “some”, “certain”) [166]. The linguistic analysis of the privacy policies revealed that those linguistic patterns are not only used by companies to avoid the continuous update of their privacy policies when new data practices are implemented. By contrast, ambiguous statements fit into a communication strategy that seeks to blur some negative data practices, precluding potential users a thorough understanding of privacy policies and, consequently, affecting the ability of individuals to provide a genuinely informed consent [161, 162].

As previously mentioned, the incompleteness of the information is another form of ambiguity that occurs in privacy policies. It shows up when the disclosed information is not able to provide adequate answers to the questions that may arise from readers, thus increasing the degree of risk perceived by individuals and decreasing their willingness to share their personal data [18]. Recent studies have addressed the problem of assessing incompleteness in privacy policy representing a statement alternatively like a frame-based structure or like an information flow, identifying a set of necessary specifications that the statement should contain, in the form of semantic roles inside the frame or parameters in the information flow [20, 177]. Those studies demonstrated that the lack of one or more necessary specification increased ambiguity and negatively affected the user comprehension of the privacy policies.

3.3 Summary

This chapter presented the legal context of the thesis with a particular focus on the GDPR, that is the framework of references that many companies inside and outside the borders of European Union must comply with, since the entry into force of the Regulation in May 2018.

Among the core concepts of the GDPR, the principle of transparency was discussed because of the resulting obligations for the data controllers, who must provide the data subjects with information about the processing activities performed on their data. This information is written in privacy policies, that are traditionally long and difficult-to-read documents, that could negatively affect a transparent communication with the data subject.

4 | Knowledge Extraction in the Data Protection Field

This chapter describes the experimental part of the thesis, implemented for detecting recurrent information scenarios from the text of the privacy policies, based on the different ontological models.

The first part of the chapter presents some NLP techniques to which the implemented experiments are related and inspired, i.e. semantic role labelling, ontology learning and open information extraction. The second part of the chapter describes an analysis of existing Ontology Design Patterns suitable to represent information scenario in the data protection domain. Based on this analysis, the third part of the chapter presents the experiments that exploited a domain-specific ontology pattern and different vocabularies to implementation the final system for detecting informative scenarios from privacy policies. Two main experiments are described: *(i)* an open information extraction task for the extraction of lexico-syntactic clues of a recurrent information in the text, *(ii)* an approach to the extraction of personal data types and purposes of the processing, driven by the concepts modelled in domain-specific and domain-independent vocabularies. Based on the results of those experiment, the last part of the chapter describes the system that integrates both the approaches in order to extract recurrent informative scenarios from the text of privacy policies.

4.1 Approaches to automated knowledge extraction from text

4.1.1 Semantic Role Labelling

Semantic Role Labelling (SRL) is the task of automatically identifying the arguments of a given predicate and assigning them semantic labels that represent the roles that those arguments play within the predicate [105]. An alternative definition has been provided by Màrquez et al. [122] who characterised the goal of the task as the detection of basic event structures such as "who" did "what" to "whom", "when" and "where". According to these definitions, the expected outcome of a SRL task would be a labelled sentence as the one in the following example:

(1) $\underbrace{\textit{Mary}}_{\text{agent}}$ *gives* $\underbrace{\textit{the present}}_{\text{object}}$ $\underbrace{\textit{to Daniel}}_{\text{recipient}}$

where the predicate *gives* represents the "what" in the event described by the sentence. The predicate is associated with different arguments corresponding to roles that represent the *agent* causing the event, the *object* directly affected by the event and the *recipient* who benefits from it.

SRL grounds on Fillmore's theory of frame semantics [74], that inspired subsequent linguistic theories and the automated approaches to solve the task. According to frame semantics, the meaning of a word is never understood on its own, but it is rather determined by the situational and experiential knowledge that the word recalls. For instance, the understanding of the verb *give*, in the previous example, is determined by the experiential knowledge we own about the "giving" action, that implies the presence of a *donor*, a *recipient* and a *theme*, i.e. the object being given. This situational and experiential knowledge can be gathered and organised in a semantic structure, i.e. the frame, made of semantic roles, referring to the specific facets that characterise the knowledge. Frames can be used to represent knowledge about actions, events and situations. A frame is always triggered by a term (not necessarily a verb), called lexical unit, that evokes the frame structure best suited for understanding the specific situation, action or event expressed by the sentence in which the lexical unit appears. In the previous example, for instance, the verb *give* is the lexical unit that actives the frame made of the three semantic roles, i.e. *donor*, *recipient* and *theme*, that contribute in the overall understanding of the sentence. The frame semantics theory proposed by Fillmore inspired the creation of FrameNet[1] [13], a semantic resource that, to date, collects 1124 frames manually labelled by a group of lexicographers.

A research related to Fillmore's theory has been developed by Levin [114], who investigated the relationship between the semantic roles evoked by a verb and the syntactic realisations of these roles in a sentence. In her research, Levin highlighted how the semantic roles associated with a verb can have multiple valid syntactic realisations (called syntactic alternations) within a sentence. For instance, the following sentence:

[1] https://framenet.icsi.berkeley.edu/fndrupal/

(2) $\underbrace{Mary}_{\text{agent}}$ *gives* $\underbrace{\textit{Daniel}}_{\text{recipient}}$ $\underbrace{\textit{the present}}_{\text{object}}$

represents the same situation described in sentence (1), showing that the verb *give* can equally express the *recipient* and *object* roles by switching the order of the direct object and the dative case. Levin studied the syntactic alternations of 3100 English verbs and grouped them in classes according to the syntactic alternations that they have in common. The research of Levin inspired the release of the PropBank, that contains all the sentences of the English Penn TreeBank annotated with semantic roles.

The divergence between the viewpoints adopted by Fillmore and Levin, being the former independent from the syntactic realisation of a role, resulted in a non standardised set of semantic roles for performing the SRL task. Indeed, the semantic roles envisioned by Fillmore and inserted in FrameNet are more specific, being tailored on the specific situation or event modelled by the frame. By contrast, the semantic roles envisioned by Levin and used in PropBank are more general and are based on the verbs classes identified in the Levin's work.

From a computational point of view, as effectively summarised by Marquez [122], the task of automatic SRL involves two main sub-tasks: *(i)* the identification of the boundaries of the arguments within a sentence and *(ii)* the labelling of those arguments according to the roles they play. Those methods usually follow a three-step processing pipeline. In the first step, the set of candidate arguments for a predicate is pruned according to some heuristic rules. In the second step, the candidate arguments are associated with a confidence score for each of the possible role labels. In the third step, the confidence scores computed independently for each arguments are combined to output a global score that represents the suitability of a given combination of labels for the arguments of a predicate.

This processing pipeline is followed by two state-of-the-art tools for SRL: SEMAFOR [47] and the Mate Semantic Role Labeler [21]. SEMAFOR is a frame-based parser that utilises a corpus of role-annotated sentences taken from both FrameNet and the SemEval 2007 data. SEMAFOR implements a three-step pipeline for extracting semantic frames from a sentence. The first step concerns the identification of target words, i.e. the lexical units that evoke frames in a sentence. Targets are detected relying on lists of seed terms that are morphological variations of the lexical units annotated in the corpus of reference. In the second steps, SEMAFOR selects the frame evoked by each target word. The system implements a discriminative probabilistic model with a latent variable for improving its performance on new target words, i.e. words that were not present in the training set. In the last step, SEMAFOR identifies the arguments for a given target associated with the corresponding frame predicted by the previous step. The association of each role of a frame with the corresponding argument in the sentence is based on the predictions of a probabilistic model whose features are derived from the sentence dependency parse tree, the words overlap between arguments and target words, and the part of speech tags of the words close to the arguments.

A three-step pipeline is also implemented by the Mate Semantic Role Labeler. In the first step, a set of local greedy classifiers is utilised to disambiguate the predicates. Afterwards, the arguments of the predicates and their corresponding role labels are identified through a beam search, resulting in a set of candidate propositions. In the second step, a global model is used to re-rank the candidate propositions, based on the local models and the propositions' features. The output of the greedy classifiers from the first step and the output of the re-ranker from the second step are finally combined, in the third step, by a linear model that finds the best candidate proposition.

A different, but still related, perspective on frame semantics was envisioned and elaborated by Mordijk, who proposed the concept of semagram [131]. From a representational point of view, a semagram is a frame-like structure made of slots and fillers. Given a semantic class of concepts, the slots represent the abstract conceptual structure shared by the concepts in the semantic class, whereas the fillers represent the actual values that the slots assume when considering a specific concept. Differently from the Fillmore's frames, however, a semagram is not intended to represent situational and experiential knowledge evoked by a word but, instead, the encyclopedic knowledge that exhaustively defines a term. Moreover, differently form Levin, the semagram structure for a semantic class is defined independently from its syntactic realisations within a sentence. During the PhD programme, I personally contributed in the development of a semi-automated approach to the creation of semagram structures for concrete concepts [113]. The approach starts from a set of hand-crafted semagrams that is extended semi-automatically with different techniques, based on the use of syntactic patterns learnt from Wikipedia and the re-use of information from existing lexical resources, i.e. WordNet and SketchEngine.

4.1.2 Ontology Learning, Population and Enrichment

The automation of the ontology building process has been foreseen as a means to overcome the knowledge acquisition bottleneck and lighten the manual workload for ontological engineers, exploiting the amount of unstructured, semi-structured and fully-structured resources available on-line [120]. Ontology learning from unstructured resources, specifically, is applied on textual corpora and it is intended as a layered framework of several tasks to be implemented in sequence, each with different methodologies and techniques that ground in fields of NLP and machine learning.

The main tasks of ontology learning, summarised by Cimiano [37], focus on the TBox component of an ontology, for automatically learning its concepts, relationships and axioms. Concepts are the first elements of an ontology to be learnt. Lexical terms are extracted from a corpus of documents that are relevant in the domain to be modelled by the ontology and groups of synonym terms are formed. Based on this terminology, the concepts of the ontology are learnt through a process of abstraction from a lexical to a semantic level. Once the ontological concepts are defined, the process focuses on learning the relationships among them. Taxonomic relations are the first to be extracted so as to define the hierarchical structure of concepts. The subsequent learning task aims to discover generic relations between concepts, exploiting their hierarchical organisation to correctly identify the domain and the range of those relations [38]. Finally, the last tasks of the ontology learning framework concern the learning of axioms and rules for providing a formal description of the classes and relationships extracted from the previous steps.

Over the years, many surveys [94, 175, 197, 10] investigated the main techniques that have been applied to address the tasks of ontology learning. Those surveys converge on the identification of three main categories of approaches, namely linguistic, statistical, and logical approaches.

The learning of ontology concepts has been traditionally addressed by statistical methods which, in their simpler version, identify as relevant those terms that appear with a high frequency in a corpus of documents. However, many approaches adopt more sophisticated variants of this metric. For instance, the Term Frequency Inverse Document Frequency (TF-IDF) metric [170] normalises the frequency of occurrence of a term in a document by the inverse of its frequency across all the documents of a corpus. The C/NC value [77], instead, incorporates contextual information to frequency, considering a fixed window of words that appear in the left and right side of a target term. In the Text2Onto framework for ontology learning [40] those metrics are combined to assess the relevance of the terms in a textual corpus, whereas Doing-Harrys et al. [60] rely on TF-IDF vectors and cosine similarity to find synonyms of seed concepts to learn an ontology for the medical field.

More recent approaches, have taken advantage of word embeddings to detect the concepts of an ontology [128, 158]. Word embedding models learn distributed representation of words from text corpora, representing each word as a real-valued vector in a predefined vector space. The benefit of including word embeddings generated by different models in an ontology learning framework was investigated in [91], while Pembeci [157] showed how ontology learning can benefit from embeddings representations also in languages other than English.

The extraction of the relations that hold among concepts is addressed by different categories of methods. Linguistic approaches are based on the analysis of the syntactic structure of sentences. In [98], Hippisley et al. exploits the syntactic modifiers of words to infer taxonomic and meronymy relations between concepts, while Sordo et al. [183] utilise dependency parsing combined with a step of Named-Entity recognition to extract relations between couples of named entities in the music field. Other branches of linguistic approaches have been based on the extraction of lexico-syntactic patterns, following the impetus given by Hearst's work [95]. For instance, Panchenko et. al. [151] inferred a hierarchical structure among concepts by combining the lexico-syntactic patterns provided by three existing systems and an approach that looks at the modifiers of words. Machine learning models have also been widely used for the extraction of relations between concepts. Hierarchical clustering, in its agglomerative variant, has been mostly utilised for learning taxonomic relations, while association rules have been applied for learning general relations between concepts, as it is made in the OntoGain system [61].

The task of learning the axioms of an ontology is usually addressed by logical approaches. Among them, Inductive Logic Programming (ILP) is an approach at the crossroad between the Machine Learning and the Logic Programming field. Given a background knowledge and a set of examples, it develops predicate descriptions in the form of logic programs. This approach has been applied, for instance, by Lehmann at al. [112] to learn formal definitions of classes in description logics and by Völker et. al [193] to learn disjointness axioms.

In the context of this thesis, which is based on the reuse of ontological and terminological resources, it is also necessary to mention the ontology learning approaches that are rooted in the use of existing computational lexicons and lightweight ontologies. Because these resources provide an easily accessible set of pre-defined concepts and relations [197], their use has been shown to be convenient in all the steps of the ontology learning process and with different types of approach. WordNet [129] is a general purpose computational lexicon that has been widely used in this context. This resource organises the terms that are linked by a synonymy relation in *synsets*, each of them representing a concept with a specific sense (i.e. its meaning), described by a natural language gloss. Moreover, synsets are linked to each other by semantic relations, like hyponymy, hyperonymy, holonyms and meronymy. The structured information provided by WordNet has been primarily exploited in the process of learning domain ontologies.

In [188], Turcato et al. learnt the concepts that are relevant in the aviation domain by implementing a statistical approach based on term frequency for finding synonyms terms. Afterwards, the synonymy relations that are found among terms are automatically validated using WordNet. Navigli et al. [139] integrated the use of WordNet in a framework for learning concepts in the tourism domain. In their approach, the glosses and the relations in WordNet are utilised to build semantic networks of words, starting from the terms extracted from a corpus of documents. Based on a weight computed for the nodes at the intersection of different networks, the words are finally associated with the most suitable WordNet sense. The tourism domain was also the focus of the framework proposed by Cimiano and Staab [39] where linguistic and statistical approaches are combined for learning concept hierarchies. In this framework, the lexico-syntactic patterns proposed in the aforementioned work by Hearst are used to extract couples of concepts where one is the hypernym of the other. These couples of concepts are jointly utilised with the taxonomic structure of WordNet to guide an agglomerative clustering algorithm. In this algorithm, two terms are put in the same cluster only if they are in an actual hyponym/hyperonym relation or if they share the same hypernym, according to the taxonomies of concepts provided by WordNet or derived from the Hearst patterns.

Ontology enrichment and ontology population are tasks related to ontology learning. Specifically, ontology enrichment refers to the task of extending the TBox component of an existing ontology with new concepts and relations, whereas ontology population refers to the task of adding new instances of concepts and relations to the ABox component of the ontology. Approaches to ontology enrichment and population can be addressed by the implementation of linguistic, statistical and machine learning methods, like those discussed for the extraction of concepts and relations for learning new ontologies [119]. Additionally, those tasks can be addressed by Open Information Extraction techniques which extract, from unstructured text, tuples of concepts linked by a relational phrase. The goals and the main techniques that have been proposed for performing Open Information Extraction are discussed in the next section.

4.1.3 Open Information Extraction

The process of converting the unstructured information embedded into a text to a set of structured facts is called Information Extraction [104]. The first approaches to Information Extraction were proposed in the early 2000s and criticised some years later by Banko et al. [14]. The researchers pointed out the lack of scalability of those approaches due to the amount of manual effort required for the definition of extraction patterns tailored to fit a specific knowledge domain and hardly adaptable to large heterogeneous text corpora. To overcome those deficiencies, they proposed a paradigm called Open Information Extraction (OIE) aimed to detect relational tuples from unstructured text, without requiring the human intervention in any step of the paradigm. Their vision was that of a domain-independent approach able to work with diverse datasets, while remaining computationally efficient.

Given an input sentence, the expected outcome of an OIE system is a tuple made of some arguments and a relational phrase that expresses the semantic association that ties the arguments. For instance, in the following sentence[2]:

Hudson was born in Hampstead, which is a suburb of London.

we would expect an OIE system to extract two tuples representing as many facts stated in the sentence:

(Hudson, was born in, Hampstead)
(Hampstead, is a suburb of, London)

where *was born in* and *is a suburb of* are the relational phrases that link the left and the right arguments in each tuple. The approaches proposed in literature to implement OIE are mainly based on the syntactic analysis of the sentence, sometimes combining lexical and semantic information to identify the relational phrases and their arguments in a sentence.

The ReVerb tool [71] extracts relational phrases that meet both a syntactic and a lexical constraint. The syntactic constraint is formulated as a regular expression on the Part of Speech (PoS) tags of a sentence. The regular expression captures relational phrases consisting either of a verb optionally followed by a preposition or a verb combined with both nouns and prepositions. The relational phrases that satisfy the syntactical constraint are further checked with respect to the lexical constraint, which is based on the assumption that a relational phrase should be general enough to extract many couples of arguments from a large corpus. Based on this assumption, the lexical constraint is verified by means of a dictionary of relational phrases that are known to satisfy the syntactic constraint and to be able to extract at least a predefined minimum number of arguments from a corpus of Web pages. Finally, a logistic regression classifier is trained on a corpus of manually labelled extractions for assigning a confidence score to the extracted facts.

[2] The example is taken from [71]

OLLIE [124] is another OIE system that relies on ReVerb for the extraction of high precision tuples where the arguments of the relational phrases are named entities. From the sentences associated to the seed tuples, OLLIE learns open pattern templates, i.e. paths that connect both the arguments and the relational words within the dependency tree of the sentence. Some of this patterns are strictly syntactic, while others may be associated with semantic and lexical constraints. When necessary, OLLIE enriches the extracted triples with extra fields for handling conditional sentences (e.g.*"If he wins five key states, Romney will be elected President."*) and non-factual sentences about beliefs or subjective remarks(e.g. *"Early astronomers believed that the earth is the centre of the universe"*).

In proposing WiSeNet, Moro and Navigli [133] embraced a more semantic perspective to OIE. Their approach is based on the remark that synonymy and polysemy are linguistic phenomena that could occur between relational phrases, enabling the identification of clusters of synonymous relations. The method is made of three steps. First, Wikipedia is used to extract relational phrases between pairs of hyperlinks in a Web page. Second, based on their similarity, the relational phases are grouped with a soft clustering technique that allows synonymous relational phrases to belong to the same cluster meanwhile allowing polysemous relational phrase to be associated with more than one cluster. The similarity between relational phrases is computed extracting the shortest dependency path that connects two arguments in a relational phrase and taking into account the distributional similarity of the words in that path. When the clusters are formed, the so-called "semantic type signatures" of synonymous relations are computed determining the Wikipedia categories associated to the arguments of the relational phrases. In the last step, the relation instances are assigned to the semantically closest cluster of synonymous relations, exploiting the semantic type signatures.

The intuitions proposed in WiSeNet have been taken up and further investigated in DefIE [56], another system that embraces a semantic approach to OIE. DefIE builds the extraction of relational phrases on the corpus of definitions in BabelNet [138], a general-purpose vocabulary that provides a semantic network of concepts linked through lexical and semantic relationships[3]. The approach performs a step of word sense disambiguation on terms and multiword expression in the definitions, linking the disambiguated phrases with the concepts of BabelNet. The relational patterns in the definitions are extracted considering the shortest paths, in the dependency graph, containing at least

[3]Because BabelNet has also been adopted in the experimental part of this thesis, the feature and the organisation of this resource will be explained more deeply in Section 4.2.1.4

one verb node and connecting disambiguated pairs of entities. The notion of "semantic type signature" of a relation, introduced in WiSeNet, is adopted also by DefIE to further specify the semantics of the extracted relations. Specifically, the signature of a relation is determined considering the hypernym common to the largest subset of arguments extracted for a relation. Finally, DefIE organises the relational phrases in a taxonomic structure that is deduced by looking for taxonomic relations between the content words that form the relational phrases.

Bucking the previously-mentioned OIE approaches that extract relational phrases containing verbs, ReNoun [198] focuses on the extraction of noun-mediated relations. For instance, from the sentence excerpt *"the CEO of Google, Sundar Pichai"*, we would like to extract the tuple *(Google, CEO, Sundar Pichai)*, where *CEO* is the noun-mediated relational phrase that links the two arguments. To do so, first, ReNoun extracts a set of seed facts for the noun-mediated relations contained in an ontology of nominal attributes (for instance, in the previous example the nominal attribute is *CEO*). Those facts are extracted with hand-crafted rules for identifying the subject and the object of the noun-mediated phrase. The parsing structure of the seed facts is used to train a distant supervision methodology that learns a set of dependency parse patterns. Finally, those patterns are used to produce new extractions from a corpus of news articles.

While all the previously mentioned approaches focus on the extraction of binary relations (i.e. relational phrases that link two arguments), KrakeN [5] tries to extract N-ary relations from sentences. The goal is to overcome the information loss due to the limitation of relational phrases to have only two arguments. For instance, in a sentence like:

> *"Doublethink, a word that was coined by Orwell in the novel 1984, describes a fictional concept"*

the relational phrase *was coined* links more than two arguments and an OIE system is expected to extract the following tuple:

> *(Doublethink, was coined, by Orwell, in the novel 1984)*

For finding the arguments of those N-ary relational phrases, the authors of KrakeN identify a set of hand-crafted paths in the dependency graph of a sentence. Some of those paths are used to extract the relational phrase from a sentence, while other dependency paths are used to identify the arguments of the relational phrase.

Similarly to KrakeN, ClausIE [55] focuses on the extraction of N-ary relations. The system assumes that a sentence can be split in minimum units of information called clauses. A clause is made of two fixed constituents, i.e. the subject (S) and the verb (V). Additional constituents of a clause may be indirect and direct objects (O), complements (C) and adverbials (A). ClausIE starts from the observation that there are only seven combinations of constituents that form a linguistically valid clause according to the English grammar. Those combinations of constituents can be used to assign a type to the clauses. For instance, the tuple *(Albert Einstein, won, the Nobel Prize)* is associated with the SVO type, being made of a subject, a verb and a direct object. The constituents that are part of a clause type are called "essential". A clause type can be, however, further extended with optional constituents, like in the clause *(Albert Einstein, won, the Nobel Prize, in 1921)*, where the basic clause type SVO is extended with an optional adverbial constituent. For extracting the clause constituents from a sentence, ClausIE relies on dependency parsing. It starts from the head verb of the sentence and detects the constituent of a clause based on the dependency tags that are found along the dependency tree of the sentence itself. Afterwards, the type of a clause is detected by some hand-crafted rules that combine the parsing information associated with the identified constituents and some properties of English verbs. ClausIE also deals with relative pronouns, possessive adjectives and participial modifiers by forming "synthetic clauses", for making explicit the relations that are not mediated by a verb. For instance, the sentence excerpt *"his discovery"* corresponds to a SVO clause *(he, "has" discovery)*, where the verb *has* is indicated between quotation marks for indicating that it has been automatically inferred from the presence of the possessive adjective *his*. Having identified the clauses types and their constituents within a sentence, the corresponding n-ary tuples (that are called *propositions* by ClausIE) are made of all the constituents, both essentials and optional, of the clauses.

All the aforementioned systems have undergone a process of manual assessment of their output. In fact, an agreed definition of the requirements that a valid tuple should meet is still missing, preventing the creation of shared annotated corpora to be used as baselines in the evaluation of the OIE systems [143]. Consequently, the common approach used to evaluate a target system is to measure the increase in precision produced by the extractions of a target system with respect to the extractions resulting from the application of other existing OIE approaches. The corpora used for the evaluation usually include a few hundred sentences, so that the task can be performed manually by two or three annotators.

Interestingly, OLLIE and ReNoun have compared their performances to those of Semantic Role Labelling approaches, among others. This comparison is based on the observation that both OIE and SRL systems focus on verb-mediated or noun-mediated propositions, making possible a comparison in their ability to detect pairs of nouns that have an asserted relationship within a sentence [124].

4.2 Automated extraction of ODPs from privacy policies

After an overview of some techniques aimed at extracting knowledge from unstructured texts, this section describes the main contribution of the thesis which investigates how ODPs can be exploited to extract and model information from the text of privacy policies. In particular, Section 4.2.1 will present different kinds of resources that have been utilised for implementing the experimental part of the thesis, discussing their scope and some limitations which have been considered when they were jointly used in the experiments. Afterwords, Section 4.2.2 will provide a preliminary overview of existing ODPs that could possibly model informative scenarios occurring in the data protection field. Finally, Section 4.2.3 and Section 4.2.4 will present two different experiments that led to the proposal of the final approach for extracting knowledge from privacy policies based on ODPs, as it will be described in Section 4.2.5.

4.2.1 Adopted Resources: Description, Scope and Limitations

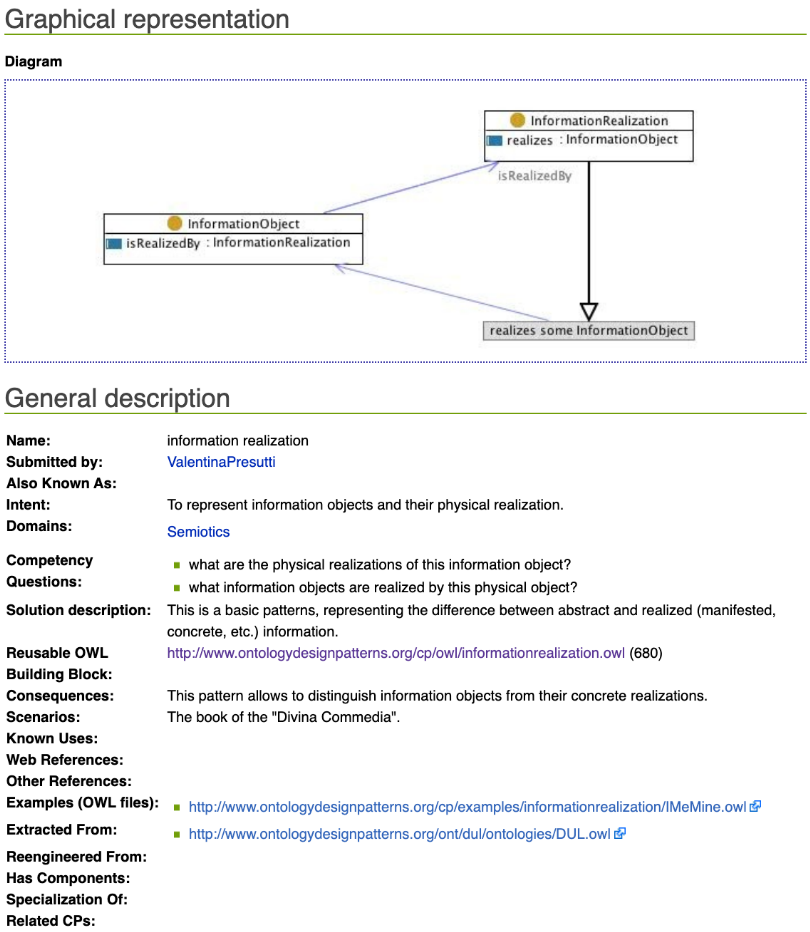
The experimental part of the thesis relied on heterogeneous resources which have been used together for implementing the extraction of information from privacy policies based on ODPs. Specifically, the *ODPs portal* has been the reference resource for discovering existing ODPs. The *OPP-115 corpus* was the set of documents adopted for performing the preliminary tests of information extraction from privacy policies, while the *Data Privacy Vocabulary* and *BabelNet* have been respectively utilised as domain-specific and general purpose vocabularies for driving the extraction of such information from the corpus.

The following subsections describe the aforementioned resources, discussing also some of their limitations and the resulting choices that were made in order jointly use them in the experimental approach.

4.2.1.1 The Ontology Design Patterns Portal

The Ontology Design Patterns portal[4] was proposed for the first time as part of the NeOn project [163]. It is intended as a virtual space for discussing about

[4] http://ontologydesignpatterns.org/wiki/Main_Page



Graphical representation

Diagram

General description

Name:	information realization
Submitted by:	ValentinaPresutti
Also Known As:	
Intent:	To represent information objects and their physical realization.
Domains:	Semiotics
Competency Questions:	▪ what are the physical realizations of this information object? ▪ what information objects are realized by this physical object?
Solution description:	This is a basic patterns, representing the difference between abstract and realized (manifested, concrete, etc.) information.
Reusable OWL Building Block:	http://www.ontologydesignpatterns.org/cp/owl/informationrealization.owl (680)
Consequences:	This pattern allows to distinguish information objects from their concrete realizations.
Scenarios:	The book of the "Divina Commedia".
Known Uses:	
Web References:	
Other References:	
Examples (OWL files):	▪ http://www.ontologydesignpatterns.org/cp/examples/informationrealization/IMeMine.owl
Extracted From:	▪ http://www.ontologydesignpatterns.org/ont/dul/ontologies/DUL.owl
Reengineered From:	
Has Components:	
Specialization Of:	
Related CPs:	

Figure 4.1: The catalogue entry for the *Information realization* CP from the Ontology Design Patterns portal: `http://ontologydesignpatterns.org/wiki/Submissions:Information_realization`.

good practices in ontology design and contributing to the enrichment of the portal by proposing new ODPs [46].

Each pattern in the portal is associated with a catalogue entry, i.e. a Web page that documents the pattern, providing all the necessary information for understanding its scope and evaluating its reuse in the process of ontology design. This information is specified following a standard template, made of several fields, including:

- the *name* of the pattern and a list of one or more *domains* it scopes;
- a description of the pattern's *intent*, i.e. the pursued representational goal, and the *competency questions* expressing the requirements the pat-

tern should fulfil;

- a description of one or more *scenarios* for providing users with actual examples that could be modelled by the proposed pattern;
- the *reusable OWL building block* that formalises the proposed pattern and a *graphical representation* that depicts its classes and properties;
- the URI of the ontology from which the pattern was *extracted* or *re-engineered*, using the referenced ontology as the base model;
- a reference to other patterns in the portal which are a *specialisation* of the proposed pattern, or which are used as its *components* or, generally, which are *related* to it because of similarities in their intent.

Because the specification of all the fields in the template that documents the pattern it not always possible, one or more fields may be left blank. Figure 4.1 shows the catalogue entry for a CP in the portal.

4.2.1.2 The OPP-115 corpus

The Online Privacy Policies - set of 115 (OPP-115) corpus [196] was released in 2016 in the context of the Usable Privacy Policy project[5], which aims to ease the approach of end users to the reading and understanding of privacy policies.

As suggested by its name, the corpus includes 115 privacy policies, issued by US-based companies. Each document in the corpus was manually annotated by three law students that followed a two layered annotation scheme, summarised in Figure 4.2. In its first layer, the scheme provides a set of ten labels to be associated to the privacy policies' paragraphs. The labels represent different data practice categories and a paragraph in a privacy policy can be annotated with zero or more labels, according to is content. In the second layer of the annotation scheme, the descriptions of the *data practice categories* are refined by a set of *attributes*, which are specific of a given data practice and which can assume a limited set of predefined *values*. Some attributes are mandatory, while other attributes are optional, depending on their relevance in defining a data practice. Moreover, each attribute-value pair can be associated with a text span in the privacy policy for providing evidence to the association between the attribute and its value.

Overall, the 3792 paragraphs of the privacy policies in the corpus are annotated with 23K data practices at paragraph level and 103K text spans are annotated with attribute-value pairs.

[5] https://www.usableprivacy.org/data

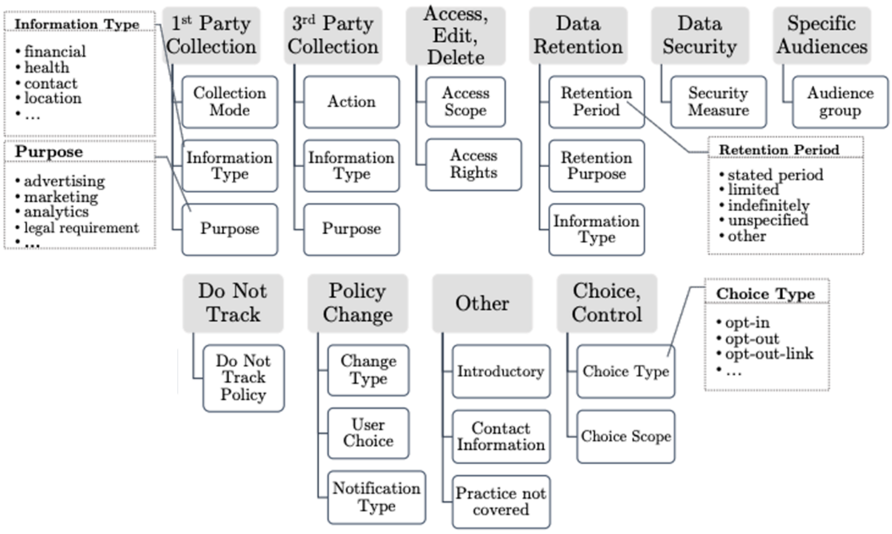


Figure 4.2: The OPP-115 annotation scheme in a graphical representation provided by Harkous et al. in [93]. The grey boxes represent the data practices, while the white boxes represent the mandatory attributes that describe each data practice. Examples of values that some attributes can assume are also provided.

As previously mentioned, the OPP-115 corpus and its annotations have been the starting point for testing the approach proposed in this thesis for extracting information from privacy policies based on existing ODPs in the portal. However, since this thesis considers the European legal framework set by the GDPR and the privacy policies in this corpus fall out of the scope of the Regulation (because they were issued by US-based companies before the entry into force of the Regulation), some limitations to the use of this corpus have been taken into account and addressed. Section 4.2.1.5 will provide a further discussion about the use of the OPP-115 corpus in the thesis.

4.2.1.3 The Data Privacy Vocabulary

The Data Privacy Vocabulary (DPV)[6] [156] has already been presented in Section 2.3 of Chapter 2, when its features were compared to those of other legal ontologies. Because the DPV has been extensively used in the experimental part of the thesis, this section provides further details about the scope and structure of this vocabulary.

The DPV considers the legal framework set by the GDPR and provides a so-called "base ontology" to model different aspects involved in the processing

[6] https://dpvcg.github.io/dpv/

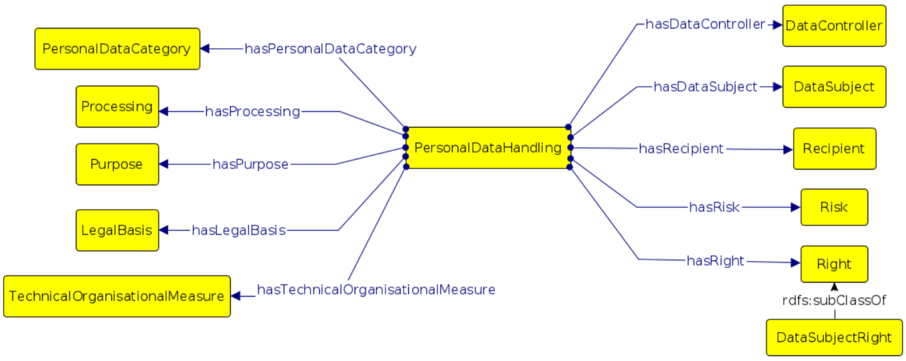


Figure 4.3: A representation of the base ontology of the DPV, as provided in its documentation Web page[6]. The *PersonalDataHandling* class is the root concept which connects the classes representing different aspects involved in a personal data processing activity.

of personal data. Those aspects include: *(i)* categories of personal data, *(ii)* purposes for processing those data, *(iii)* legal grounds the processing is based on, *(iv)* categories of operations performed in the processing activity, *(v)* legal entities involved, *(vi)* technical and organisational measures implemented for a secure processing, *(vii)* rights applicable in the data processing activity and *(viii)* possible risks involved. Each aspect is modelled by a different class in the base ontology and it is linked to the other aspects through the *Personal Data Handling* concept, as shown in Figure 4.3.

Most of the concepts in the base ontology are further specialised by dedicated sub-vocabularies which provide the DPV with a modular organisation. Each module (i.e. sub-vocabulary) consists of a taxonomy of concepts, linked by the `rdfs:subClassOf` property of the RDFS data model. Not all the taxonomies in the respective modules have the same level of detail along the vocabulary. For instance, the taxonomy that models the categories of personal data is the most developed and reaches a high level of specificity. Indeed, considering the taxonomic organisation of the concepts as a tree-like structure, the maximum depth of the nodes[7] in that module is equal to four. An example of the taxonomy of concepts for modelling categories of personal data is provided in Figure 4.4.

Similar taxonomies, but with a lower depth, are also provided for the other aspects involved in the processing of personal data. Only the concepts of *Right* and *Risk* in the base ontology lack a deeper taxonomic organisation. Specifi-

[7]When referring to the depth of a node, I consider the definition provided in `http://typeocaml.com/2014/11/26/height-depth-and-level-of-a-tree/`: "The depth of a node is the number of edges from the node to the tree's root node". According to this definition, the most general class in the taxonomy lies at depth 0, all its direct subclasses lie at depth 1, and so on.

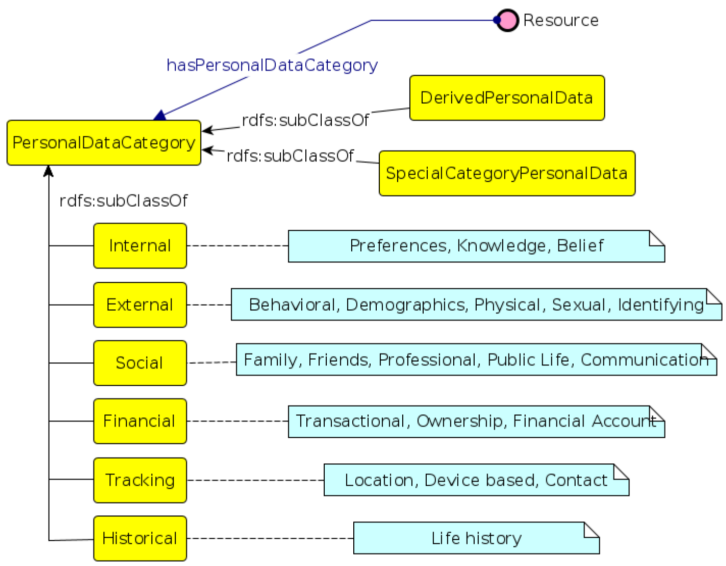


Figure 4.4: Some of the concepts in the *Personal Data Category* module, in a graphic representation provided in the DPV documentation Web page[6]. The *PersonalDataCategory* concept is the root of the taxonomy and it is also included in the base ontology (see Figure 4.3). The list of concepts in this Figure is not exhaustive. In particular, the concepts listed in the blue squares are just a few examples of the concepts that are modelled at different depths in the taxonomy.

cally, the *Right* concept is only specialised by one further class (i.e. the *DataSubjectRight* class) and the *Risk* concept has not subclasses that specialise it. One possible explanation for that lack of refinement could be that the *Right* and *Risk* concepts have been introduced only in the last version of the vocabulary, released in January 2021. Since the DPV is an ongoing work, some enrichment for those concepts could be expected in future releases.

Other further considerations about the vocabulary and the concepts modelled in it will be discussed in the next sections, in connection with the discussion of the experimental setting for the extraction of information from privacy policies.

4.2.1.4 BabelNet

BebelNet[8] [138] is a multilingual knowledge base that was automatically built integrating a semantic lexicon for English, i.e. WordNet[9] [129], and the knowledge extracted from a well known collaborative online encyclopedia, i.e. Wikipedia[10].

[8] https://babelnet.org/
[9] https://wordnet.princeton.edu/
[10] https://en.wikipedia.org/wiki/Main_Page

The core of BabelNet consists of a direct graph having the word senses in WordNet as vertices and the semantic relations connecting those concepts as labelled directed edges. Each vertex of the graph is called *Babel synset* and it represents a word sense, i.e. a meaning. Similarly to WordNet, each Babel synset is associated with its lexicalisations, which are synonym terms that express that meaning. Through an automatic mapping between the word senses (i.e. the vertices of the graph) and the titles of the Wikipedia pages, the semantic graph was enriched with unlabelled edges that represent unspecified relations drawn from the different types of hyperlink connecting the Wikipedia pages. Inter-language links among Wikipedia pages, specifically, were used to find lexicalisations of a word sense in different natural languages, relying on statistical machine translation systems when those links where not available.

Since 2012, the BabelNet knowledge graph has been keeping growing and it automatically integrates, in its current version[11], the knowledge extracted from ten additional resources. BabelNet offers Java and Python API to access the resource programmatically as well as a SPARQL endpoint to formulate query over its RDF graph.

4.2.1.5 Scope and Differences of the Resources

The resources described in the previous subsections are different in several ways. They vary in format, year of publication and scope, and those differences must be taken into account when planning to jointly use them.

The Ontology Design Patterns portal and BabelNet are two domain independent resources. They do not refer to a single field of knowledge or reality, but rather they adopt a broader view on different areas of knowledge. On the one hand, the portal does not set constraints on the types of scenarios that a submitted pattern can model, on the other hand, the cross-domain extent of BabelNet lies at the core of its knowledge graph, built on two comprehensive resources like WordNet and Wikipedia. Consequently, the adoption of these two resources to represent knowledge and to automatically process texts in the law filed must be properly balanced. Indeed, as already mentioned in other parts of this thesis (see Chapter 1 and Chapter 3), the law field has several peculiarities, like the intentional ambiguity of legal texts and the multitude of interpretations arising from them. A single general purpose resource is unlikely to capture and correctly represent the complexity and variety of the legal domain, consequently, its applicability should be subject to precise constraints and assessments.

[11] In April 2021, the current version of BabelNet is 5.0

By contrast, the OPP-115 corpus and the DPV are domain-dependent resources that refer to the privacy field. Despite this commonality, however, there are notable differences between them. First, they differ in their format which, in turn, depends on the objective for which they were released. The OPP-115 corpus is made of a set of CSV files (one for each privacy policy) containing the two-layered annotations. The corpus is mainly meant for training machine learning models which perform text classification tasks, as it is shown in the corpus reference paper (see the above-mentioned reference [196]) where the resource is used to train three different classifiers for predicting the paragraph-level labels. By contrast, the DPV is released as an RDF file, available in different serialisations. The adoption of Semantic Web standards makes the DPV suitable both to promote the interoperability between compliance checking tools and to enable its extension according to specific users' needs. Furthermore, while the DPV provides a taxonomic organisation of its terminology using the `rdfs:subClassOf` property, a similar hierarchical structure lacks in the labels of the OPP-115 corpus. Despite some works have referred to the corpus' annotation scheme as a *taxonomy*[12], there is not a formal structuring of the associations between the attribute-value pairs used in the annotation of the text spans and the paragraph-level labels.

The second difference between the resources concerns the legal framework which they refer to. The DPV models the data protection domain as it is framed by the GDPR that applies to companies and entities that are established in the EU as well as companies that, despite being established outside the EU, address their goods/services to the EU citizens (see Chapter 3). By contrast, the OPP-115 corpus collects privacy policies that were issued by US companies some years before the entry into force of the GDPR, consequently the Regulation does not apply on them and some concepts modelled in the DPV can not be expected to be mentioned in the privacy policies of the OPP-115 corpus. For instance, in the DPV, the module that represents the legal grounds for a fair processing of personal data (see the representation of the DPV base ontology, in Figure 4.3) refers to a specific requirement set by the GDPR, but the privacy policies of the OPP-115 corpus will not clearly express this information that, manifestly, lacks in the annotation scheme, both at paragraph and text span levels (see the graphical representation of the annotation scheme in Figure 4.2).

The OPP-115 corpus seems an outdated resource, considering the increasing attention that is being devoted, not only in Europe, to the drafting of legal

[12]Works like [93] and [116] refer to the annotation scheme as *Wilson taxonomy*, from the name of the first author of the reference paper for the OPP-115 corpus, i.e.[196]

frameworks for ensuring the right to privacy and data protection[13]. However, the corpus is still used, to date, in many approaches that have applied NLP techniques to process the text of privacy policies (see Chapter 6 for an overview of state-of-the-art approaches). The success in the adoption of this corpus is mainly due to its free availability online. Moreover, a similar open-access collection of annotated privacy policies released after the entry into force of the GDPR, and to which the framework set by the Regulation applies, does not exist yet. The lack of such a resource was also witnessed in 2019 by Gallé et al. [78], where the authors envision the possible extensions, targeting the GDPR, of the annotation schemes of two existing corpora of privacy policies, one being the OPP-115 corpus and the other being a corpus of privacy policies referring to the GDPR framework, but limited in size and not released for open access.

Having acknowledged the lack of appropriate corpora framed in the GDPR context, I chose to rely on the OPP-115 corpus to perform the experiments aimed to extract, from privacy policies, the information that is relevant with respect to some existing ODPs. This choice was also made with the aim of exploiting the annotations provided by the corpus as a means of carrying out a step-by-step assessment of the pipeline that has been implemented and which will be explained in the following sections. Thereby, it was possible to implement a semi-automatic evaluation of the intermediate results without requiring a “human in the loop”. In fact, although a manual assessment of all the intermediate outcomes of the pipeline would have been more reliable and influential, it would have required the involvement of one or more legal experts, remunerating them for the time spent in accomplishing the assigned evaluation tasks.

When the OPP-115 corpus was jointly used with the DPV, the experiments were constrained to take into account the differences underpinning the two resources. Moreover, to reconcile the experiments and the results obtained on the OPP-115 corpus with the GDPR framework, a final evaluation of the proposed approach was performed on a set of privacy policies in the GDPR scope.Those documents were manually selected from a larger corpus of privacy polices, i.e. the Princeton-Leuven longitudinal corpus which, being first released in March 2020, was not available when the first experiments were performed. The details about this corpus, the selection of the privacy policies in the GDPR scope and the performed evaluation will be presented in Chapter 5.

Considering the growing attention to the field of personal data protection and the in-development, but still scarce, landscape of corpora and vocabular-

[13] An example of this trend is the California Consumer Privacy Act [11] which became effective on January 1, 2020.

ies embracing an EU perspective on the field, the aim of the performed experiments was to rely as much as possible on the existing available resources. Consequently, the following sections will represent an effort of analysis, harmonisation and integration of existing information sources, with the aim of filling the gap between the unstructured text of privacy policies with more formal patterns of knowledge, represented by existing ODPs. The content of the next sections will be articulated in an initial analysis of the ODPs portal for identifying the patterns that are potentially suitable for the legal field. Then, the focus will shift on a set of experiments aimed to apply some NLP techniques for mapping the unstructured information in privacy policies on a selected ontological pattern for representing information in privacy policies, taking also advantage of the DPV and BabelNet for driving the selection of the information.

4.2.2 ODPs for the Data Protection Domain: a Preliminary Selection

As discussed in Chapter 2, ODPs promote standardised representations of the common modelling scenarios occurring in a domain of interest. In this section, I present an analysis of the ODPs that are listed in the ODPs portal (see Section 4.2.1.1). This analysis aims to clarify the extent to which existing ODPs address modelling scenarios which could occur when the domain of interest concerns the data protection field. Because I was interested in finding ODPs that *conceptually* model the domain of interest, the proposed analysis focuses on CPs, listed in a dedicated Web page of the portal[14]. By contrast, I did not consider other families of ODPs which address expressivity and architectural problems (as mentioned in Section 2.2.3) out of the scope of the thesis.

There are 163 CPs submitted in the ODPs portal. To find those patterns that could be of interest to the data protection field, the analysis involved an iterative process of subsequent eliminations of not-relevant patterns. At different stages of the process, exclusion criteria are applied to the CPs in order to discard those that do not meet them. The choice to rely on a "backward" process, working by exclusion, was deemed as the most practical solution for the selection of the patterns. A "forward" procedure would have required a two-step analysis made of: *(i)* an *a priori* elicitation of all the possible modelling scenarios that could occur in the data protection field and *(ii)* a search in the portal for existing CPs corresponding to the elicited scenarios. That procedure would have been a highly time-consuming activity, asking for the involvement of several legal advisers with expertise both in the data protection and in the knowledge representation fields. Moreover, it would hardly have been possible to exhaustively list all the possible modelling scenarios, as they

[14] http://ontologydesignpatterns.org/wiki/Submissions:ContentOPs

Iteration	Elimination Criterion	Removed Patterns
1st	The content pattern lacks of the competency questions or the OWL building block.	69
2nd	The content pattern is associated with a domain label that is weakly related to the data protection domain.	20
3rd	The content pattern is associated with a domain label that is likely to be related to the data protection domain, but the competency questions reveal a poor correlation to the domain of interest.	12
4th	The content pattern lacks of a domain label and the competency questions reveal a poor correlation to the domain of interest.	20

Table 4.2: The criteria applied to delete noisy CPs from the list provided by the portal. Each row specifies the elimination criterion applied in a specific iteration of the process and the number of patterns that, meeting that criterion, were excluded to the list of patterns of interest to the data protection field.

depend on the specific modelling requirements of a project. By contrast, this backward analysis allows the identification of a list of pre-selected CPs, cleared of "noisy" patterns, both for the incompleteness of their documentation and their low relevance with respect to the domain of interest. Room is left, then, to the possibility to further refine this analysis based on the specific modelling requirements of possibly interested users.

The proposed process is made of four iterations that were performed manually. Table 4.2 summarises the criteria that I set to filter out the noisy patterns in each iteration.

The first iteration aimed at removing the CPs that are missing an appropriate documentation. Some CPs, indeed, are associated with a catalogue entry where one or more fields of the documentation template are left blank. Specifically, I discarded the patterns lacking of the competency questions or the OWL building block. The former information is important for framing the modelling scenario addressed by the pattern, whereas the latter is necessary to formalise the pattern in a machine-readable and reusable representation. If this type of information is missing, the pattern evaluation process is compromised, hindering an assessment of the pattern relevance in the domain of interest. This intuition is consistent with the results provided by some surveys [118, 92] that investigated the significance of different fields in the documentation template for ascertain the reusability of a pattern. The performance of the first iteration of the process lead to the discarding of 69 CPs from the portal, with 94 CPs left for the following steps of the process. The high number of deleted CPs (representing the 41,7% of the total number of CPs) reveals a significant weakness in many CPs submitted in the portal. Indeed, by not providing an appropriate documentation, those patterns preclude possible interested users from reusing them.

After discarding patterns missing basic documentation, the second and the third iterations of the process focused on the *domain* field of the documentation template. The portal collects 68 labels that refer to as many domains with different levels of specificity. Some labels evoke highly specialised domains, like *Fishery*, *Biology* and *Internet of Things*, while other labels denote more general topics, like *Time*, *Planning* and *Systems*. Moreover, the portal provides a *General* label for those patterns which are not specialised or limited to a range of subjects. Some CPs are associated with one or more labels that represent the domains of interest they refer to, whereas other CPs are not associated with any specific domain label. To the best of my knowledge, there is not a specific reason which explains why some CPs are missing domain labels. Instead, it seems that the choice not to specify a domain for a pattern is at the sole discretion of the user who submits a new CP to the portal.

Based on the labels associated with the *domain* field, the filtering criterion in the second iteration of the analysis aimed to exclude those patterns referring to a domain which is less likely to be related to the data protection field. Some examples of those domains are *Fishery* and *Biology*, but also *Smart City*, *Agriculture* and *Physics*. Taking into account the insights provided by the aforementioned surveys [118, 92], which indicates the *domain* field as scarcely relevant for the assessment of a pattern, the application of this exclusion criterion has been as conservative as possible. Only those domains for which the non-relation to the privacy domain was most clear have been deleted, so as to postpone to the next iteration a deeper analysis of the patterns whose relevance for the domain of interest was uncertain. The application of the exclusion criterion in the second iteration of the process resulted in the deletion of an additional 20 CPs.

The patterns that were preserved by the second iteration went through the third phase of the process, where I read their competency questions and inspected their OWL building blocks in order to clarify their relatedness to the data protection field. I deleted 12 additional patterns in this third step, providing, for each of them, an explanation for their discarding.

The forth and last iteration of the process involved the remaining 62 patterns, being those that lack a domain label. Similarly to the third iteration, I analysed their competency questions and their OWL building blocks to unravel those modelling scenarios which have a weak correlation with the data protection field, providing an explanation for their exclusion. This step of analysis led to the discarding of further 20 CPs. Appendix A contains the list of patterns that were excluded at each iteration of the process.

The remaining 42 CPs were those that passed all four elimination steps of the process. The use cases they model have been assessed to be of possible interest to the data protection field, and the documentation they provide should allow an assessment of their suitability with respect to specific modelling requirements. Figure 4.5 shows those patterns, grouped by their domain labels[15]. Looking at the Figure, it is clear how the set of CPs unravels the multi-faceted nature of the data protection field, that could involve a variety of heterogeneous scenarios to be considered. When it comes to the modelling of the data protection field, possible use cases could concern (but not be limited to)

- the modelling of agents who play different roles like, for instance, the data controller or the data processor roles;
- the modelling of personal data flows, as the GDPR requires data controllers to maintain a record of the data processing activities under their responsibility (Art. 30);
- the modelling of events happening at a specific point in time, like, for instance, a data breach event.

The domain labels and the names associated to the selected CPs evoke this multitude of scenarios. Moreover, a cross-domain look at the uses cases modelled by the patterns can reveal further similarities among groups of patterns. For instance, some CPs model use cases that involve the presence of an agent (intended as a human being). This is the case of the *Acting For*, *Agent Role*, *Part Of* and the *Participation* patterns, to name a few. Other patterns represent actions and events that require the modelling of temporal parameters. This is the case, for instance, of the *Activity Specification*, *Action* and *Time Indexed Participation* patterns.

[15] To ease the visualisation of Figure 4.5, I did some simplifications in representing the domain labels. Specifically: *(i)* I omitted to represent the association of a pattern with the *General* label when that pattern was also associated with some other, more specific, domain labels; *(ii)* I represented the *Management* and the *Scheduling* domains as a single domain, because they always occur together in the list of selected content patterns.

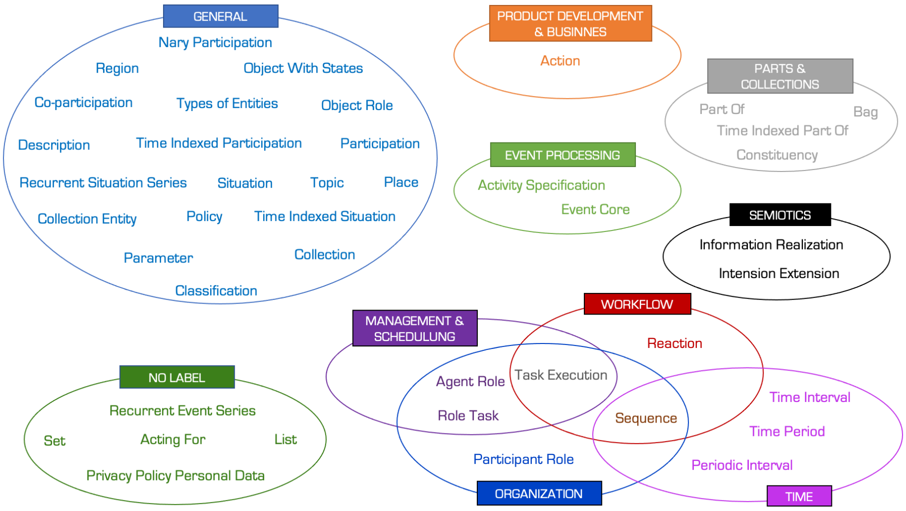


Figure 4.5: Distribution of the forty selected CPs over their domain labels, including the patterns that were missing the specification of a label and that are represented, in the figure, as *no label*.

Among the CPs that passed the four-step elimination process, the *Privacy Policy Personal Data* pattern is of particular relevance to the data protection domain. According to the documentation provided by the corresponding catalogue entry in the ODPs portal, the intent of the pattern is to represent the information disclosed within a privacy policy [155], referring to the data protection framework set by the GDPR in the modelling solution it provides. Indeed, the proposed pattern reuses some of the concepts defined in the GDPRtEXT [153] and GDPRov [154] ontologies, that were described in Section 2.3.1.4 of Chapter 2 and that specifically tackle the modelling of concepts related to the Regulation.

Given the appropriateness of the *Privacy Policy Personal Data* pattern to model a use case in the domain of interest, the approach that I will discuss, in the next sections, for the extraction and the representation of knowledge from privacy policies was partially guided by the concepts modelled by this pattern. In particular, the next section will present a preliminary study for the retrieval of textual evidence of the scenario modelled by this pattern in the privacy policies of the OPP-115 corpus.

4.2.3 Identification of Recurrent Text in Privacy Policies

4.2.3.1 Introduction and Premises

The analysis performed in the previous section identified the *Privacy Policy Personal Data* pattern as one of the relevant CPs in the data protection field. Because the declared intent of the pattern is to model the information within a privacy policy, the preliminary experiment presented in this section aimed to detect the presence of the modelling scenario expressed by the pattern in the text of privacy policies.

The assumption underlying this experiment was that, if a CP should represent a recurrent ontology design problem, then evidence of this recurrence could be retrieved in the texts belonging to the domain of interest. The type of evidence looked for in the privacy policies is made of lexico-syntactic clues that unravel the presence of the information modelled by CP. Ideally, these patterns should be derived from the syntactic structure of a sentence, combining information about the presence of certain words within the sentence with the information about their syntactic role in it.

The extraction of the lexico-syntactic clues was envisaged as an Open Information Extraction task on the privacy policies of the OPP-115 corpus. The decision to rely on OIE was made after reading the text of some privacy policies, both from the OPP-115 corpus and from on-line Web services. The text of those privacy policies usually adopts the form of a "one-way dialoge" between the company and the data subject. In this dialogue, the company describes the data practices performed on personal data by addressing the data subject explicitly. For instance, in a sentence like:

> *"We [...] collect some information from your computer or device automatically as you use our service."*

the company (Skyscanner[16], in this example) explains a data practice performed on personal data by using the *we* pronoun for referring to itself as the entity that performs the processing activity and using the pronoun *you* for addressing the data subject. In this dialogue-like communication, the company states in first plural person the performed processing activities, expressed by verbs (*collect*, in the sentence above). This communicative style was found in many privacy policies, both in the OPP-115 corpus and from on-line Web services.

Based on these empirical observations, I assumed that an OIE tool could extract relational phrases containing the verbs that express different processing

[16] https://www.skyscanner.ie/media/privacy-policy

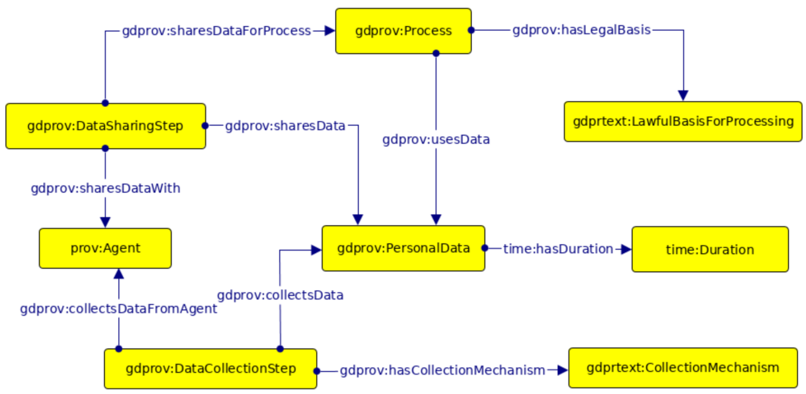


Figure 4.6: The *Privacy Policy Personal Data* pattern, in the visualisation provided by the ODPs portal.

activities described in the privacy policies. Moreover, because the description of the processing activities performed on personal data is the target of a privacy policy for guaranteeing the data subject her right to receive information (see Section 3.1.2.5), I expected relational phrases about processing activities to be extracted with high frequencies from those documents.

4.2.3.2 Experimental Setting for the Open Information Extraction Task

To verify the previous intuitions, I chose ClausIE (see Section 4.1.3) as the tool for performing the OIE task. I deemed this tool as the most suitable for performing the task because of the explicit association that it makes between the constituents of the clauses extracted from a sentence and the syntactic roles they play in it. In fact, I assumed that this information could have supported the verification of the remarks that I made about the communicative style of the privacy policies, which was based on the analysis of some syntactic elements of their sentences.

The *Privacy Policy Personal Data* pattern, shown in Figure 4.6 models different aspects involved in the personal data processing. Consequently, before the application of the OIE tool, I identified smaller groups of classes within the pattern. To determine those groups, I considered one class at a time and I identified as a group the set made of that class and all the classes in the range of a relationship having the class under consideration as a domain. What I found is that each group represents a particular sub-scenario of the personal data handling macro-scenario modelled by the pattern. The identified groups refer to:

(i) the collection of personal data performed by the company, including the

Group Identifier	Group of Classes in the CP	OPP-115 Data Practice
(i)	DataCollectionStep, PersonalData, Agent, CollectionMechanism	*First Party Collection/Use*
(ii)	DataSharingStep, Personal Data, Process, Agent	*Third Party Sharing/Collection*
(iii)	PersonalData, Duration	*Data Retention*
(iv)	Process, PersonalData, LawfulBasisForProcessing	*First Party Collection/Use,* *Third Party Sharing/Collection,* *Data Retention*

Table 4.3: The groups of subclasses that were identified within the *Privacy Policy Personal Data* pattern and the corresponding data practices, from the OPP-115 corpus, that were associated to the groups. The roman numerals in the first column refer to the description of each group of subclasses provided in Section 4.2.3.2

.

types of personal data collected, the agent from which those data are collected and whether the personal data are obtained by automated means;

(ii) the activity of sharing personal data performed by the company, including the types of shared data, the recipients of those data and the purpose of the data sharing;

(iii) the retention period of the personal data;

(iv) the purposes and the legal grounds that apply on different processing activities performed on personal data.

I also exploited the labels provided by the paragraph-level annotation scheme of the OPP-115 corpus (see Section 4.2.1.2) to define some correspondences between the annotated data practices in the corpus and the groups of classes that I identified in the pattern. Table 4.3 shows those correspondences, that were defined after reading the descriptions of the data practices provided by the OPP-115 corpus documentation and the competency questions associated to the pattern. The first three groups of classes are associated with the *First Party Collection/Use*, *Third Party Sharing/Collection* and *Data Retention* data practice labels, respectively. The forth group of classes, instead, is associated with all three of the aforementioned data practices, because the GDPR requires each data processing activity to be motivated by a purpose that is legitimated by a legal ground. As already mentioned when discussing the limitations of the adopted resources (see Section 4.2.1), the proposed correspondences must be intended on a coarse level and some aspects modelled in the pattern (i.e. *LawfulBasisForProcessing*) are not applicable to the OPP-115 corpus, due to the different legal frameworks referred by these resources. This preliminary experiment, indeed, aimed to find lexico-syntactic patterns that reveal the presence of the scenario modelled by the pattern, according to the groups of subclasses identified, without establishing precise mappings between the text and the pattern's classes.

The ClausIE tool was applied on the paragraphs of the OPP-115 corpus that were labelled with one of the data practices in the third column of Table 4.3. From the output of ClausIE, I considered the SVO clauses, i.e. those containing a subject (S), a verb (V) and an object (O). The number of SVO clauses extracted from the paragraphs associated with the same label in the corpus was high, especially for the paragraphs labelled as *First Party Collection/Use* and *Third Party Sharing/Collection*, from which 3820 and 3296 clauses where extracted, respectively. The clauses extracted from the paragraphs labelled as *Data Retention* were only 670, that is consistent with the smaller number of annotations referring to this data practice in the corpus.

To analyse the results, I ordered by decreasing frequency the clauses extracted in the paragraphs having the same data practice label. I focused on the 50 most frequent clauses extracted for each data practice to see if any of them might reveal a particular aspect of the data handling scenario identified by one of the four groups of classes in the pattern. For each aspect, I selected five clauses that I judged representative of it, as shown in Table 4.4. Some considerations about the outcome of the OIE task are provided in the following section.

4.2.3.3 Insights from the Results of the Task.

From the exemplary clauses included in Table 4.4, some considerations can be made with regard to the assumptions formulated before the execution of the OIE task and described in Section 4.2.3.1.

The first consideration refers to the dialogue-like structure and the communicative style that was noticed by reading some privacy policies. The extracted clauses, regardless the data practice they refer to, highlight this communicative style. Indeed, most of them show the pronoun *we* in the constituent that plays the role of subject in the clause. Another evidence of this communicative-style comes from those clauses showing the verb *has* between quotation marks. These are the "synthetic" clauses that ClausIE automatically builds to deal with

Clauses for Group (i)	freq.
<we, collect information>	276
<you, “has", name>	161
<you, “has", address>	122
<you, provide, information>	82
<we, receive, information>	54

Clauses for Group (ii)	freq.
<we, share, information>	223
<we, disclose, information>	114
<you, “has", name>	91
<we, provide, information>	35
<we, sell, information>	33

Clauses for Group (iii)	freq.
<you, “has", account>	39
<we, retain, information>	21
<we, store, information>	17
<you, “has", name>	12
<we, delete, information>	10

Clauses for Group (iv)	freq.
<we, use, information>	345
<you, “has", name>	161
<you, “has", address>	209
<we, use, address>	29
<we, combine, information>	25

Table 4.4: Some of the clauses extracted from the paragraphs of the OPP-115 corpus. The clauses were associated to a specific aspect of the the data handling scenario modelled by the *Privacy Policy Personal Data* pattern. The roman numerals identify the groups of classes represented in Table 4.3.

the possessive adjectives. For instance, from the textual excerpt *“your name”*, the tool infers a clause having the pronoun *you* as the subject and *to have* as verb. As Table 4.4 shows, those clauses frequently occur in all the considered sub-scenarios modelled by the pattern, confirming the communicative style of the privacy policies, where the data subject is the direct addressee of the communication.

A second consideration from the obtained results emerges by looking at the verbs in the clauses. In the top-50 frequent clauses extracted from the paragraphs of each data practice, I found verbs that are representative of the particular processing activity performed on personal data. For instance, the verbs *collect* and *receive* suggest the presence of an information related to the collection of personal data, consequently the corresponding clauses were associated to the first group of classes, (i.e. group *(i)*). Similarly, verbs like *share* and *disclose* recall a personal data sharing activity (i.e. group *(ii)*), whereas the verbs *retain* and *store* evoke a data retention practice (i.e. group *(iii)*). The clauses containing these verbs appear in privacy policies as introductory statements before explaining the details of a data practice, that usually mentions the personal data involved in the processing activity. Sentences within the privacy policies are structured in a way that is similar to the following one:

> *“When you register, we may collect personally identifying information, including your name [...]”*[17].

for which ClausIE extracts, among others, the clauses *(we, collect, information)*

[17] This sentence excerpt is taken from the *Meredith* privacy policy, in the OPP-115 corpus.

and *(you, "has", name)*. This example also suggests an explanation for the high extraction frequency of "synthetic triples", resulting from the presence of possessive adjectives associated with terms that recall personal data.

Similar observations can also be made for the forth, more general, group of classes identified in the pattern (i.e. group *(iv)*). The verb *use* in the clause *(we, use, information)* was found in the top-50 frequent extractions from the paragraphs of each data practice. This occurrence across different data practices is demonstrated by the high frequency associated to it in Table 4.4. This clause is extracted from sentences like:

> *"We use the information we learn from you to help us personalize and continually improve your experience on the Sites"*[18].

where the sentence excerpt *"we use the information"* introduces a statement concerning the purpose of the processing activity, being a personalisation service in this example.

Overall, the performance of the OIE task on the text of the privacy policies confirmed the intuitions formulated before its execution. The text of the privacy policies shows commonalities across documents, regarding the way the sentences are formulated. The clauses extracted from the documents represent the lexico-syntactic patterns whose presence was assumed when the task was conceived. The high frequency of synthetic clauses containing a reference to a personal data in their object constituent was an unexpected result, but one that further revealed how certain lexico-syntactic patterns may function as clues, within the text, of an information relevant to the pattern's modelling scenario.

4.2.4 Vocabulary-driven Extraction of Concepts from Privacy Policies

4.2.4.1 Introduction and Premises

In Section 4.2.3, I focused on the extraction of recurrent clauses from the text of privacy policies, for finding lexico-syntactic clues that evoke the presence of a processing activity, as modelled by the *Privacy Policy Personal Data* CP. In this section, instead, I present the approach that I implemented for finding more specific information about the processing activities described in privacy policies. For the implementation of this step, I relied on the DPV that was first presented in Chapter 2, and further discussed in this chapter (see Section 4.2.1.3). As already noticed in Section 2.3.1.4, the DPV vocabulary is slightly different from the other existing legal ontologies released to model the privacy

[18]The sentence excerpt is taken from the *The Atlantic* privacy policy, in the OPP-115 corpus.

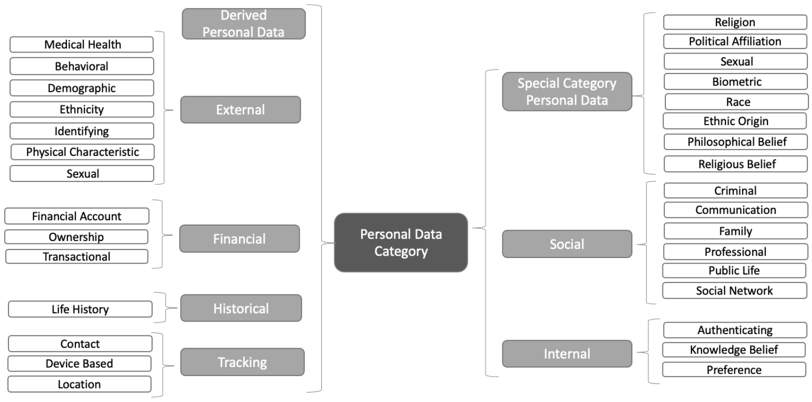


Figure 4.7: A graphical representation of the taxonomic structure of concepts in the *Personal Data Category* module in the DPV. The figure only shows part of the concepts of the taxonomy, i.e. those at a depth less than or equal to two. Most of the concepts in this figure are also represented in Figure 4.4. However, here I try to provide a clear representation of the concepts at a depth equal to two in the taxonomy, as it could not be clear enough in Figure 4.4

field. Indeed, while the other resources presented in that section rely on a higher level of formalism, the light-weight structure of the DPV provides a set of taxonomies to organise the concepts related to the data protection field. Consequently, the intuition was to rely on this vocabulary as a terminological source to guide the extraction, from privacy policies, of text excerpts related to some of the modules included in the vocabulary.

Similarly to the OIE task explained in the previous section, I used the OPP-115 as the corpus of reference for testing the proposed approach. To take into account the differences between the DPV and the OPP-115 corpus, concerning their underlying legal frameworks (see Section 4.2.1.5), some constraints were set to narrow the focus of the experiment to a limited number of modules in the DPV and data practices in the corpus. Indeed, the guided extraction of text excerpts that are relevant with respect to the DPV concepts is limited to the types of personal data collected by the company and the purpose for which such data are processed. This information is represented by the *Personal Data Category* and *Purpose* modules of the DPV, whose taxonomic structure is partially represented in Figure 4.7 and Figure 4.8. Similarly, I only considered the paragraphs of the OPP-115 privacy policies that were assigned to the *First Party Collection/Use* label in the corpus, as I expected this information to be more likely to be found within them. This assumption is also supported by the results obtained from the OIE task presented in the previous section. The "synthetic" clauses containing a reference to personal data and the clauses that refer to the purpose of the processing activity (group *(iv)* in Table 4.4) were

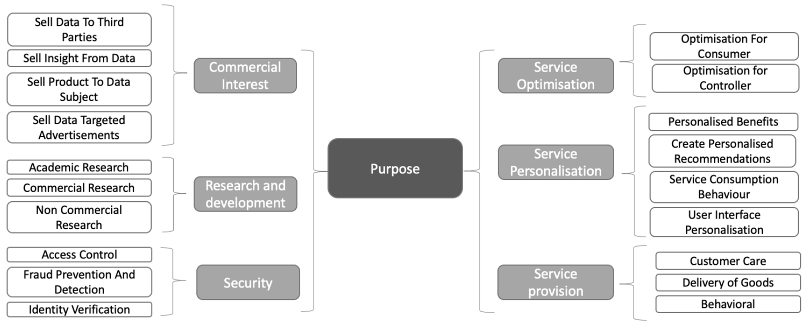


Figure 4.8: A graphical representation of the taxonomic structure of concepts in the *Purpose* module in the DPV. The figure only shows part of the concepts of the taxonomy, i.e. those at a depth less than or equal to two.

extracted with the highest frequencies from the paragraphs referring to the *First Party Collection/Use* data practice.

An effort has also been made to enrich and extend the information in the DPV with that provided by BabelNet (presented in Section 4.2.1.4). The choice of adopting a general-purpose vocabulary in a domain-specific task is justified by the communication target of the privacy policies. Those documents are used to describe the data processing activities implemented by companies that offer goods and services in a variety of areas of the everyday life. The terms used in privacy policies to describe the data processing activities are expected to span this multitude of areas. Consequently, the adoption of a general purpose vocabulary may help in the retrieval of additional concepts that are related to the DPV, but missing from it.

The next subsections will describe how the DPV and BabelNet were jointly used to guide the process of extraction of personal data categories and purposes of the data processing. The method is made of three sequential steps, where the output of one step becomes the input of the following one. The first step creates broad mappings between some parts of the text in the privacy policies and the modules of the DPV. The second step tries to refine these mappings selecting, from the modules in the DPV, some classes that could be suitable for the refinement. The last step chooses, from the set of suitable classes, the one that will yield the needed refinement.

4.2.4.2 Broad Mappings of Text Chunks on the DPV modules

The first step of the implemented method aimed to discover the parts of the text in privacy policies that are relevant with respect to the DPV. For identifying these parts of the text, I exploited the distinctiveness of the terminology

DPV Module	Top-6 of the frequent words
Purpose	(service, 17), (user, 9), (product, 8), (research, 8), (optimisation, 7), (datum, 6), (activity, 6), (commercial, 6), (recommendation, 6), (interface, 4), (individual, 4), (purpose, 4)
Personal Data Category	(individual, 148), (information, 141), (history, 18), (health, 17), (personal, 17), (social, 13), (credit, 13), (datum, 13), (professional, 11)

Table 4.6: The six most frequent words used to name and describe the classes in the *Purpose* and *Personal Data Category* modules of the DPV. The number next to each noun represents the frequency of the noun in the module. More than six terms are present in both lists due to the tie in the frequencies.

that characterises each module of the vocabulary. This evidence was found collecting and ordering, by decreasing frequency, the terms used to name the classes in each module and to provide the description of their meaning in natural language (through the RDF property `dct:description`). As Table 4.6 shows, the collected terms are in most of the cases exclusive for each module and only few words overlap. Thus, the nouns in each list can be considered as *descriptors* for the type of information that each module of the DPV represents. In this approach for detecting the relevant terminology in each module, the number of descriptors depends on the richness of the taxonomies and by the exhaustiveness of the concepts' descriptions in the modules. To mitigate this hurdle, I enriched the lists of descriptors by considering their synonyms, that were automatically retrieved from BabelNet and were considered as having the same frequency value of the descriptor to which they refer.

The lists of descriptors were used to select sentence excerpts to be mapped on the DPV modules. For each sentence, the noun chunks (i.e. the nominal phrases) were extracted using the available libraries of the SpaCy dependency parser[19]. The chunks roots (i.e. the words connecting the noun chunks to the rest of the parsed sentence) were utilised to assign the chucks to one of the two modules of interest in the DPV. When the root of a chunk matched a descriptor, the chunk was mapped on the corresponding module. In case of a match with a descriptor that appears in both the modules, the chunk was assigned to the module where the descriptor has the highest frequency. In case of a tie, the chunk was preliminarily assigned to both modules. The chunks whose roots did not find a match with a descriptor were considered as not relevant in establishing a match with the DPV. Two examples of the mappings

[19] `https://spacy.io/`

performed in this step are shown below. The module assigned to each chunk is indicated in a square box and the roots of the chunks, used to determine the mappings, are underlined.

$\boxed{\text{Purpose}}$ *customer service* $\underset{\text{root}}{\underline{\textit{purpose}}}$

$\boxed{\text{Personal Data Category}}$ *mobile device unique id* $\underset{\text{root}}{\underline{\textit{number}}}$

4.2.4.3 Detection of Candidate Classes for the Refinement of the Broad Mappings

Having as input the coarse assignments of noun chunks to one or two modules in the DPV, the second step focused on the refinement of these mappings identifying a set of more specific candidate classes in the taxonomies of the modules.

Given a text chunk, a first control checks if the name of a class in the DPV, or one of its synonyms retrieved from BabelNet, matches the chunk or appears as a sub-string in it. If this is the case, then the set of candidate classes is made of a single element, i.e. the matching class. For instance, the fragment considered in the previous Section:

$\boxed{\text{Purpose}}$ $\underset{\text{dpv:CustomerCare}}{\underline{\textit{customer service}}}$ $\underset{\text{root}}{\underline{\textit{purpose}}}$

contains the sub-string *customer service* that is a synonym of the string *customer care*. In turn, the latter matches the homonym DPV concept (see Figure 4.8, where the *Customer Care* concept is a subclass of *Service Provision*), that is considered as the candidate class to refine the mapping with the *Purpose* module.

If no class is detected with this first check, then the lists of modules descriptors (see Section 4.2.4.2) are used to find a set of candidate classes. Specifically, for each descriptor that matches a word in the text chunk, the class from which the descriptor was extracted is added to the list of candidate classes. If a candidate class is a leaf in the taxonomy of a module, then it is substituted by its direct superclass in order to avoid matches with too specific classes. The root of the text chunk is excluded in this search for candidate classes, because it already contributed to the broad mappings with the DPV modules. For instance, in the following fragment:

$\boxed{\text{Personal Data Category}}$ *mobile* $\underset{\text{dpv:DeviceBased}}{\underline{\textit{device}}}$ $\underset{\text{dpv:Identifying}}{\underline{\textit{unique}}}$ *id* $\underset{\text{root}}{\underline{\textit{number}}}$

the word *device* matches the homonym descriptor, derived from the description of the *Device Based* class in the *Personal Data Category* module (see Figure 4.7, where the *Device Based* concept is a subclass of *Tracking*). Similarly, the word *unique* matches a descriptor derived from the *UID* (i.e. user identifier) class in the *Personal Data Category* module. However, as the *UID* class is a leaf in the taxonomy of the module, its direct superclass, i.e. *Identifying*, is added to the set of candidate classes of the chunk (see Figure 4.7, where the *Identifying* concept is a subclass of *External*).

4.2.4.4 Selection of the Class for Refining the Broad Mappings

The third and last step of the method selects, among the candidate classes, the most suitable for refining the broad mapping between a text chunk and a module. The class is selected by computing the cosine similarity between the text chunk and its candidate classes. The vector representation of both the text chunk and its candidate classes was obtained from the pre-trained GloVe word embeddings[20] [158] that were combined according to some weights for representing the different contributions given by each word in the overall vector representation. When deciding about the pre-trained words embeddings to utilise in this experiments, I also considered the Law2Vec embeddings, specifically generated from a corpus of legal documents [34]. However, during the implementation of the vocabulary-driven approach, I found that many embeddings for domain-independent words (like *IP address* or *e-mail address*) were missing. This lack is explainable, once more, with the communicative style of the privacy policies and their reference to concepts spanning domains other than the legal one. Among the domain-independent words embeddings, I chose between Word2Vec and GloVe solutions, deciding for the latter because the corpus on which the embeddings model was trained is more varied, including both Wikipedia pages and new articles, as opposed to Word2Vec word embeddings, that were trained on a corpus of news articles only.

Having the GloVe embeddings, the vector representation for a text chunk is obtained collecting the set W_F of the embeddings for the content words in the chunk and the set W_S of the embeddings for the content words that occur in the same sentence of the text chunk. Assuming that all the words in the chunk contribute equally to its vector representation, a weight equal to 1 is assigned to each word embedding in W_F. By contrast, the weights associated to the word embeddings in W_S assume that the contribution of a word occurring in the same sentence of the chunk is equal to the frequency of that word in the sentence divided by the total number of distinct words in the sentence. The vector representation of the text chunk is computed, then, by multiplying each embedding in the set W_F and W_S by the corresponding weight and computing the mean vector resulting from the two sets.

By contrast, the vector representation for a candidate class is conceptually based on the computation of some Term Frequency-Inverse Document Fre-

[20] `https://nlp.stanford.edu/projects/glove/`. I used the 300-dimensional vectors.

	Purpose	Personal Data Category	Total
Chunks (with repetitions)	852	4025	4877
Chunks (no repetitions)	224	747	971
Retrieved classes	17	85	102

Table 4.7: Statistics about the number of text chunks that were retrieved in the privacy policies and the number of classes of the DPV that were associated with at least a text chunk.

quency (TF-IDF) scores associated with the words used in its description. The TF-IDF measure [170], in its basic form, normalises the frequency of occurrence of a term in a document by the inverse of its frequency across all the documents of a corpus. By doing this, terms that frequently appear in every document of the corpus, as it is the case with stopwords, will result in a law TF-IDF value. By contrast, terms that appear in a document, but are not frequent in the rest of the corpus, will result in a higher TF-IDF value.

Following the assumption that underpins the TF-IDF measure, terms that are used in one or few class descriptions should be emphasised, because they likely are more representative of a specific DPV class, whereas terms that are used frequently in the definitions of the classes should have less relevance. Therefore, being C the set of candidate classes for a text chunk, the TF-IDF scores for the content words used in the description of a class c in C were computed considering the frequency of these words in the description of c and the inverse document frequency of these words with respect to the definition of the other classes in C.

The embeddings for the content words in the description of c were then multiplied by the corresponding TF-IDF scores and the average vector of the embeddings was computed to obtain the vector representation of c.

Finally, the cosine similarity is computed between the vectorial representations of the chunk and of each candidate class. The class that results in the highest cosine similarity value is considered as the best candidate for the refinement. The example below shows the similarity values computed for the text chunks discussed in the previous sections (the class that determined the final mapping is highlighted in bold).

Purpose *customer service* (**dpv:CustomerCare** **0.77**) *purpose* (root)

Personal Data Category *mobile* *device* (**dpv:DeviceBased** **0.76**) *unique* ($dpv{:}Identifying$ 0.60) *id* *number* (root)

DPV concept	Text spans found with BabelNet
Health Record	"medical history"
UID (i.e. user identifier)	"id", "identification number"
Username	"screen name", "login name"
Location	"geographical location", "geographic location"
IP Address	"internet protocol address", "internet address"
Customer Care	"customer service"

Table 4.9: Some of the text spans that were associated with the corresponding concept in the DPV relying on the synonymy relations expressed in the BabelNet vocabulary.

DPV concept	Text spans retrieved based on the *Purpose* descriptors
Create Personalized Recommendations	"personalised ad", "personalised promotion", "recommendation service"
Research and Development	"product development purpose", "research analysis", "research purpose"
Service Provision	"health care service"
Service Personalisation	"user authentication purpose"

Table 4.10: Some examples of text chunks extracted relying on the descriptors for the *Purpose* module.

4.2.4.5 Automated Evaluation of the Detected Mappings

Statistics about the Performed Mappings. Table 4.7 shows a summary of the number of text chunks that were extracted with the methodology described in the previous sections. Overall, 4877 chunks were extracted from the privacy policies of the corpus and they were associated to 102 classes of the DPV (out of a total of 192 classes in the two modules of interest). Each chunk occurs one or more times in the corpus of privacy policies.

Omitting the repetitions, the number of unique text chunks that were retrieved is equal to 971. Among them, 128 chunks were detected because the name of a class, or one of its synonyms in BabelNet, matched the chunk or appeared as a sub-string in it. Specifically, the text chunks extracted based on a synonymy relation in BalbelNet were 43. Table 4.9 shows some of the text chunks that were mapped on the classes of the DPV based on BabelNet.

Among the 971 unique chunks, the remaining 843 chunks were retrieved populating the lists of candidate classes, relying on the descriptors extracted for each module. Table 4.10 and Table 4.11 provide some example of the mapping established for the *Purpose* and the *Personal Data Category* modules based on their descriptors.

DPV concepts	Text spans retrieved based on the *Personal Data Category* descriptors
Behavioral	"browsing habit", "click stream data"
Device Based	"unique device identifier", "device unique advertising identifier"
Identifying	"unique application number", "unique numerical identifier"
Professional	"employment information", "job information"

Table 4.11: Some examples of text chunks extracted relying on the descriptors for the *Personal Data Category* module.

Attribute Values	Classes in the *Personal Data Category* Module
Financial	Financial [1]
Health	Medical Health [2]
Contact	Contact [2], Name[3]
Location	Location [2]
Demographic	Demographic [2], Physical Characteristic [2], Professional [2], Family [2]
Personal identifier	Identifying [2], Financial Account[2]
User online activities	Behavioral [2], Social Media Communication [3]
User profile	Identifying[2], Preference[2]
Social media data	Social Network [2]
IP address device ids	Device Based [2]
Computer information	Device Based [2]
Cookies tracking elements, Survey data, Generic personal information, Other, Unspecified	

Table 4.12: Correspondences between the values of the *Personal Information Type* attribute in the OPP-115 corpus and the classes in the *Personal Data Category* DPV module. The last row lists the attribute values that did not find a match in the module.

Precision Assessment of the Performed Mappings. The evaluation of the results relied on the annotations of the privacy polices provided by the OPP-115 corpus. To estimate the precision of the mappings extracted by the implemented method, I created a correspondence between the values of the *Personal Information Type* attribute of the OPP-115 corpus and some of the DPV classes in the *Personal Data Category* module. Those correspondences were manually identified analysing the descriptions provided both for the attribute values in the corpus and the classes in the DPV, unravelling similarities in the type of information that they represent. Table 4.12 shows the mappings that I considered. In this table, the numbers between squared brackets represent the depth of a class in the taxonomy of the module (see Footnote 7 for the definition of "depth"). Most of the correspondences were made between an attribute value and a class at depth 2 in the module. I found that some attribute values are very general and no meaningful correspondences could be established. A similar analysis was also performed on the values of the *Purpose* attribute in the OPP-115 corpus and the classes of the homonym module in the DPV. Table 4.13 shows the mappings that I considered. In this case, most of the attribute values

Attribute Values	Classes in the *Purpose* Module
Basic service/feature	Service Provision [1]
Additional service/feature	Service Provision [1], Service Personalization [1]
Advertising	Service Personalization [1]
Marketing	Commercial Interest [1], Service Personalization [1]
Analytics/research	Research And Development [1], Service Optimization [1]
Personalisation/Customisation	Service Personalization [1]
Service Operation Security	Security [1]
Legal Requirement, Merger/Acquisition, Other, Unspecified	

Table 4.13: Correspondences between the values of the *Purpose* attribute in the OPP-115 corpus and the classes in the *Purpose* DPV module. The last row lists the attribute values that did not find a match in the module.

	Purpose	Personal Data Category	Total
Match	114 (13.4%)	1351 (33.6%)	1465 (30.0%)
Mismatch	296 (34.7%)	858 (21.3%)	1154 (23.7%)
No annotation	442 (51.9%)	1816 (45,1%)	2258 (46.3%)

Table 4.14: Results of the evaluation that is based on the manual drawing of the correspondences between attribute values in the OPP-115 corpus and the classes in the DPV according to the three different scenarios discussed in Section 4.2.4.5. Percentages are computed with respect to the total number of noun chunks extracted for the corresponding module.

were associated with classes at depth 1 in the *Purpose* module.

Based on the drawn correspondences, I identified three different scenarios for the evaluation. Given a text chunk f that is extracted from a sentence s in a privacy policy: *(i)* f is part of a text span in s and the attribute-value pair associated to the span matches the class of f, following the correspondences that were identified for the evaluation; *(ii)* f is part of a text span that is labelled in s, but the attribute-value pair associated to the span does not match the class associated with f; *(iii)* f does not correspond to any of the text spans that were annotated in s. The number of text chunks that fit each of the three scenarios is shown in Table 4.14. Some insights from the evaluation are presented in the next section.

4.2.4.6 Insights from the Results of the Evaluation

The first insight that comes from the retrieved mappings concerns the coverage of the two modules of interest in the DPV with respect to the classes that were associated with some text chunks in the privacy policies (see the last row of Table 4.7). The number of classes that were automatically mapped on the text chunks slightly exceeds (53.1%) half of the concepts represented in the DPV modules of interest. However, it should be noticed that many concepts in

	Personal Data Category	Purpose
Most Frequent	(Device Based, 758), (Email Address, 282), (Contact, 183)	(Commercial Interest, 337), (Purpose, 266), (Security, 49)
Less Frequent	(Philosophical Belief, 1), (Disciplinary Action, 1), (Thought, 1)	(Access Control, 1), (Service Optimization, 1), (Optimisation For Consumer, 7)

Table 4.15: DPV classes with the highest and lowest number of text chunks mapped on them.

the DPV are very specific and likely difficult to find in the privacy policies text. Classes like *Music* or *Accent* in the *Personal Data Category* module were not mapped on any text chunk. By contrast, chunks related to the *IP Address, Location* and *Contact* classes were frequently extracted. This intuition is reinforced by looking at Table 4.15, that provides an excerpt of the classes for which the highest and lowest number of text chunks (considering repetitions) was found.

Concerning the evaluation technique explained in Section 4.2.4.5, I noticed that most of the labels mismatches occurred because the text spans in the corpus were associated to general labels (like *Other*). In this case, the vocabulary-driven approach could provide an advantage over the manual annotation proposed in the corpus, suggesting more precise labels for the text spans. By contrast, the scenario in which a text chunk, that was automatically extracted by the method, but was not annotated in the corpus, needs further investigations for evaluating to what extent the lack of an annotation indicates an incorrect automatic mapping or is rather a corpus fault.

Another insight specifically concerns the performance of the system in the extraction of the purposes of the data processing. This information is not always expressed by a single noun chunk. Instead, purposes of the processing could be expressed by more articulated verbal phrases, like in the following sentence:

> *"We use to recommend features, products and services that might be of interest to you."*

In this example, the vocabulary-driven approach, that solely analyses the noun chunks, fails to detect a statement about the purpose of the processing activity, because it is expressed by the verbal chunk *"to recommend features, products and services*". This limitation of the approach could explain its low performance in extracting concepts in the DPV *Purpose* module, as it is shown in Table 4.14.

I took into account and tried to overcome this limitation by integrating the output of this vocabulary-driven extraction with the information about the syntactic structure of a sentence extracted by ClausIE, which can identify both verbal and noun chunks. However, before describing how the output of the OIE task and the vocabulary-driven approach where integrated, the following subsection will present a possible representation of the outcomes of the vocabulary-driven approach exploiting an existing ODP.

4.2.4.7 Semantic Web Oriented Representation of the Results

In this section, I propose a machine-readable representation of the mappings that were automatically extracted by the vocabulary-driven approach to concepts extraction. The understanding that I propose about the mappings detected by the system is that of *semantic domains* that are identified by the concepts of the DPV, and *domain elements* that correspond to the text chunks and that are related to the semantic domains. A standardised modelling solution to this intuition was looked for in the list of pre-selected ODPs that resulted by the analysis of the ODPs portal, as described in Section 4.2.2. I found that such a modelling solution is provided by the *Collection* Ontology Design Pattern (ODP)[21], that represents the membership of an item to a domain, not to be intended in the sharp sense defined in the set theory (as specified by the documentation provided for the ODP).

I used the RDF syntax to formalise the mappings extracted from the privacy policies by using the representational model provided by this ODP. For each DPV class that was associated with a text chunk in a privacy policy, a new class representing a related semantic domain was introduced. The text chunks were then associated to their semantic domains with the property `isMemberOf`, introduced by the ODP. The properties `skos:label` and `skos:example` were used to associate the chunks with their natural language strings and the sentences of the privacy policy from which they were extracted, as shown in the example below.

```
:DemographicDomain rdf:type dpv:Demographic, owl:Thing.

:DemographicAnalysisConcept rdf:type skos:Concept, owl:Thing;
  odp:isMemberOf :DemographicDomain;
  rdfs:label "demographic analysis"@en;
  skos:example "Perform statistical, demographic, and marketing
  analyses of users of the Sites and their purchasing patterns"@en.
```

This example shows the advantage of the proposed representation. An unstructured delivery of the results could erroneously suggest that, if the concept *Demographic*, that represents a type of personal data, contributed to the identification of some text chunks, then those chunks should only refer to other related personal data. From this sharp viewpoint, the mapping of the *Demographic* concept with the *demographic analysis* chunk would be considered incorrect, because it does not refer to a personal data type.

[21] http://ontologydesignpatterns.org/wiki/Submissions:Collection

By contrast, the representation of a semantic domain related to the *Demographic* concept and the association of the text chunk with this domain provides a new perspective on the proposed mapping. Indeed, *demographic analysis* and *demographic personal data* are different in their meaning, but it is likely that a demographic analysis will involve the processing of demographic personal data, thus legitimating a mapping of the text chunk with the corresponding domain.

4.2.5 An Integrated Approach for the Detection of Recurrent Scenarios in Privacy Policies

4.2.5.1 Introduction and Premises

The OIE task and the vocabulary-driven detection of concepts from privacy policies showed that some correspondences could be drawn between the information in those documents and the existing resources that address the conceptualisation of the data protection field from multiple perspectives. On the one hand, the OIE task identified lexico-syntactic patterns, expressed in the form of clauses, that act as clues of an information referring to the modelling scenario of the *Privacy Policy Personal Data* pattern. On the other hand, the vocabulary-driven concepts extraction found mentions to personal data categories and purposes of the processing referring to the concepts modelled in the DPV modules.

The method that I present in this section joins the findings from both the experiments to further specialise the mappings between the textual information in privacy policies and the conceptual model provided by the *Privacy Policy Personal Data* pattern. The objective is to detect the sentences that refer to data processing activities and establish mappings between the details about those processing activities and the corresponding concepts in the ontology pattern.

While the OIE task and the vocabulary-driven concepts detection were tested separately on the OPP-115 corpus, on which the US privacy regulatory framework applies (see Section 4.2.1.5), the execution and the evaluation of the method described below did not rely on that corpus. Instead, I used a set of privacy policies on which the legal framework set by the GDPR applies. The objective was to verify that the assumptions and the findings in the previous experiments are still valid in the text of privacy policies that were written for complying with the Regulation. This choice required the collection of an *ad-hoc* corpus of privacy policies as well the definition of an annotation task for the manual evaluation of the results, that will be presented in the next chapter. Meanwhile, this section describes the implementation details for combining

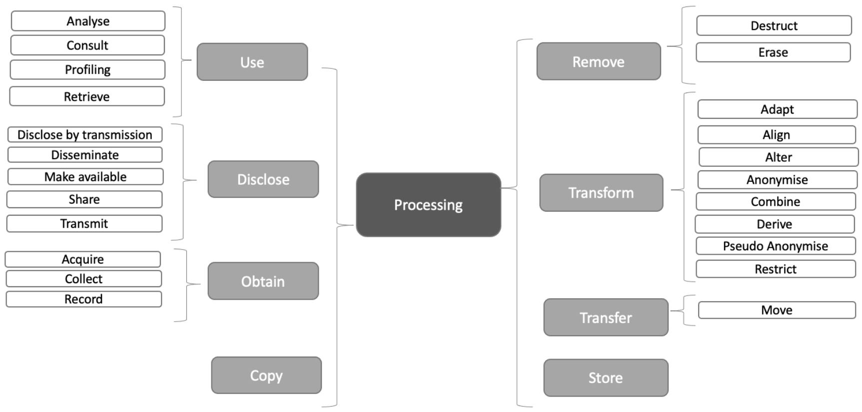


Figure 4.9: The taxonomic structure of the *Processing* module. The Figure shows all the concepts of the module.

the output of the OIE task and the vocabulary-driven extraction of concepts for finding the details related to the data processing scenarios described in privacy policies.

The method is made of two steps. The first step detects, within a privacy policy, the sentences that describe a processing activity, based on the clauses extracted from ClausIE and the concepts in the *Processing* module of the DPV. The second step finds in those sentences the details about the described processing activities, based on the conceptual model provided by the *Privacy Policy Personal Data* pattern and combining information about the clauses in the sentences and the output of the vocabulary-driven concepts detection.

4.2.5.2 Detection of Data Processing Activities from Clauses and the DPV Concepts

The information provided by a privacy policy does not only pertain to the processing activities performed on personal data. According to Art. 13 and Art. 14 of the GDPR (see Section 3.1.2.5), there are other categories of information that should be provided to the data subject, such as the rights she can exercise, the contact details of the data controller and whether her personal data will be transferred to third countries or international organisations. This information is not addressed by the modelling scenario in the *Privacy Policy Personal Data* pattern, consequently, the first step of the method focused on selecting the sentences of a privacy policy that mention a processing activity.

For implementing this step I leveraged the findings of the OIE task, expanding to a further module the application of the DPV for the extraction of information from privacy policies. The *Processing* module, indeed, provides a taxonomy of concepts, expressed by verbs, for representing the processing activities that could be performed on personal data, as shown in Figure 4.9. The method utilises those concepts to select the sentences from which ClausIE extracts an SVO clause having a verb that lexically matches the name of a class in the *Processing* module. This implementation choice was made consistently with the insights drawn from the analysis of the output of the OIE task, when I observed that some of the most frequent SVO clauses, extracted form the OPP-115 corpus, contain a verb that evokes a processing activity (see Section 4.2.3.3). However, I did not consider the whole taxonomy of concepts in the module, but I narrowed the focus on some of them, taking into account the mappings between frequent lexico-symtactic clues and different processing aspects that were found performing the OIE task. Specifically, I identified the following correspondences between some concepts in the *Processing* module, the lexico-syntactic clues and the processing aspects that were represented in Table 4.4:

- *Obtain*: its description in the DPV refers to the processing activities for gathering personal data. Among its subclasses, it includes the class *Collect*, that appears as a verb in one of the SVO clauses that were considered as lexico-syntactic clues for a data collection activity.

- *Disclose*: its description in the DPV refers to the activity of disclosing personal data. Its subclass *Share* was found as a verb in one of the SVO clauses that were considered as a lexico-syntactic clue for a data sharing activity;

- *Store*: its description in the DPV refers to the practice of keeping data for future use. This description is consistent with the information about the retention of personal data modelled by the pattern, as also witnessed by the presence of the verb *store* in one of the clauses that was frequently extracted from the privacy policies and associated with a data retention activity;

- *Use*: the description of the concept provided by the DPV is generic, as it only claims: *to use data*. This verb was extracted frequently as a constituent of an SVO clause deemed as lexico-syntactic clue for the elicitation of the purpose of the processing.

The correspondences between the identified concepts in the DPV, considering also their subclasses, and the processing scenarios modelled by the pattern, are summarised in Table 4.16.

Group Identifier	Group of Classes in the CP	DPV concepts
(i)	DataCollectionStep, PersonalData, Agent, CollectionMechanism	***Obtain*** - *Acquire, Collect, Record*
(ii)	DataSharingStep, Personal Data, Process, Agent	***Disclose*** - *DiscloseByTransmission, Disseminate, MakeAvailable, Share, Transmit*
(iii)	PersonalData, Duration	***Store***
(iv)	Process, PersonalData, LawfulBasisForProcessing	***Use*** - *Analyse, Consult, Profiling, Retrieve*

Table 4.16: The correspondences identified between the processing scenarios modelled in the *Privacy Policy Personal Data* pattern and the concepts in the DPV. The group identifiers and the groups of classes in the leftmost columns correspond to those identified in the OIE task and presented in Table 4.3. The third column lists the corresponding concepts identified in the *Processing* module of the DPV. In each row, the first concept (in bold) is the most general of the list, while the other concepts are its subclasses.

Based on those correspondences, the method applies the ClausIE tool on the text of the privacy policies. For each sentence, it checks if one of the extracted SVO clauses contains a verb that lexically matches a class in the DPV *Processing* module, according to the mappings represented in Table 4.16. Moreover, based on the intuition about the communicative style of privacy policies, the method additionally checks if the subject of the clause corresponds to the pronoun *we*. An example is provided by the sentence below:

> *"We $\underbrace{\text{share}}_{\textbf{dpv:Share}}$ your information with our service providers for the purpose of receiving services such as security, fraud detection and prevention, reporting, and storage."*

From this sentence, ClausIE extracts, among the others, the SVO clause (*we, share, information*), where the pronoun *we* is the subject of the clause and the verb *share* matches one of the DPV concepts that were identified for expressing a data sharing activity. Consequently, the sentence is selected by the method and passed to the next level of processing, that will automatically detect details about the described processing activity.

4.2.5.3 Characterisation of the Processing Activities

Having identified the sentences that refer to a data processing activity, based on the *Processing* module of the DPV, the second step aims to find details about those activities. The goal is to provide a more complete characterisation of the

DPV concepts	Optional constituents and their frequencies
Acquire	(A: to, 2), (A: however, 1), (A: include, 1), (A: occur, 1), (A: often, 1), (XCOMP: expand, 1)
Analyse	*not found*
Collect	**(A: from, 220)**, (A: about, 109), (A: also, 88), (A: in, 75), (A: knowingly, 74), **(A: for, 57)**, (A: under, 55), **(A: automatically, 45)**, (A: through, 39)
Consult	(A: have, 2), (A: prior, 2), (A: before, 1), **(A: for, 1)**, (A: provide, 1), (A: visit, 1)
Disclose	**(A: to, 146)**, (A: in, 69), (A: also, 38), **(A: for, 29)**, (A: how, 28), (A: only, 23), (A: without, 20), (A: with, 18), (A: require, 16)
Disseminate	*not found*
MakeAvailable	*not found*
Obtain	**(A: from, 21)**, (A: about, 15), (A: so, 12), (A: also, 8), **(A: for, 8)**,(A: in, 8), (A: through, 8), (A: before, 6), (A: with, 6)
Profiling	*not found*
Record	**(A: automatically, 7)**, (A: by, 3), **(A: for, 3)**, (CCOMP: recognize, 3), (A: about, 2), (A: also, 2), (A: as, 2), (A: from, 2), (A: independently, 2)
Retrieve	(A: from, 3), (A: by, 2), (A: on, 2), (CCOMP: help, 2), (A: for, 1), (A: return, 1)
Share	**(A: with, 376)**, (A: in, 55), (A: also, 46), **(A: for, 45)**, (A: about, 33), (A: how, 29), (A: provide, 15), (A: as, 11), (A: describe, 11)
Store	(A: in, 23), (A: on, 20), (A: also, 14), (A: automatically, 13), **(A: for, 12)**, (A: choose, 8), (A: from, 8), (A: visit, 8), (A: elect, 6)
Transmit	**(A: to, 22)**, (A: of, 5), (A: during, 4), (A: in, 4), (A: use, 3), (A: also, 2), (A: pay, 2), (A: typically, 2), (XCOMP: use, 2)
Use	(A: also, 225), (A: how, 204), **(A: for, 169)**, (A: in, 125), (A: when, 111), (**XCOMP**: collect, 65), (A: with, 63), (**XCOMP**: help, 63), (**XCOMP**: provide, 61)

Table 4.17: The concepts of the DPV *Processing* module and their most frequent optional constituents, extracted by ClausIE when those concepts appear as verbs in the SVO clauses. The optional constituents that were considered for finding details about the processing activities are highlighted in bold.

processing scenarios described by the sentences, mapping specific text excerpts to the concepts in the *Privacy Policy Personal Data* pattern.

As mentioned in Section 4.1.3, ClausIE distinguishes between the essential and the optional constituents of a clause. In the previous step, the essential constituents of the SVO clauses were used to find the sentences that describe a data processing activity. In this step the analysis of those clauses is further enlarged to their optional constituents, in order to characterise those activities with more detail. Moreover, in performing this step, the vocabulary-driven approach for concepts extraction is applied to the sentences and its the output is combined, when possible, with the information about the clauses that are part of those sentences.

One of the cases in which the analysis of the optional constituents proves useful concerns the adverbials in the SVO clauses. Depending on the processing activity evoked by the verb of the clause, those adverbials are introduced by a preposition that suggests the semantic role played by an agent in that activity. Looking at Table 4.16, the *Agent* concept appears in the groups of classes that were identified for referring to activities of data collection and sharing. The set of verbs that are used in the DPV to model those activities allows a limited set of adverbials to be used for representing an agent who provides or receives personal data. On the one hand, the agent from which personal data are obtained is expected to appear within an adverbial introduced by the preposition *from*. On the other hand, the agent to whom personal data are disclosed is expected to appear in an adverbial introduced by the prepositions *with* or *to*, depending on the specific verb in the clause. This intuition can also be verified considering the clauses extracted from the sentences in the OPP-115 corpus and looking at the frequency with which a verb in an SVO triple is associated with several optional constituents, as shown in Table 4.17. As it can be noticed, the adverbials (indicated by the letter *A*) introduced by the identified prepositions frequently co-occur with the verbs evoking a processing activity. Consequently, if such adverbials appear as optional arguments in the SVO clauses , then the system relies on this information to identify the agents involved in the processing activity. For instance, from the sentence that was introduced in the previous section:

> *"We* $\underset{\textit{\textbf{dpv:Share}}}{\underline{\textit{share}}}$ *your information* $\underset{\textit{\textbf{odp:Agent}}}{\underline{\textit{with our service providers}}}$ *for the purpose of receiving services such as security, fraud detection and prevention, reporting, and storage."*

the ClausIE tool extracts the clause *(S:we, V:share, O:information, A*:with)*, where the optional adverbial is represented in the clause by the preposition *with*. The corresponding proposition is underlined in the sentence and it is associated with the *odp:Agent* label, where the prefix *odp* indicates that the class *Agent* belongs to the *Privacy Policy Personal Data* pattern.

The information about the agent involved in a processing activity is not the only type of detail that can be detected analysing the optional constituents of a clause. In the *Privacy Policy Personal Data* pattern, a data collection scenario is characterised by the class *Collection Mechanism*, that specifies whether or not the data are gathered by automated means. By looking at an adverbial constituent introduced by the word *automatically*, the method can identify those data collection activities that are explicitly stated to be automated. This adverbial, indeed, was frequently found within those clauses where the verb refers to a data collection activity. (see Table 4.17).

For extracting the purposes of a data processing activity, the main intuition was to exploit the output of the vocabulary-driven concepts extraction, specifically relying on the mappings found between noun chunks in the sentence and concepts in the DPV *Purpose* module. Indeed, the information modelled by that module corresponds to the same information which is modelled, in turn, by the *Processing* class in the *Privacy Policy Personal Data* pattern. Moreover, the optional constituents of the SVO clauses are utilised in this step to overcome some of the limitations of the vocabulary-driven concepts extraction. Indeed, as discussed in Section 4.2.4.6, the purpose of a processing activity could be expressed within a sentence by verbal phrases that are more articulated than the noun chunks extracted by the implemented vocabulary-driven extraction. Consequently, when the approach detects a noun chunk representing a purpose, the whole proposition, extracted by ClausIE, in which the chunk appears is used to represent the purpose of the processing. For instance, in the sample sentence mentioned above:

> *“We* $\underset{\textbf{dpv:Share}}{\underline{\textit{share}}}$ *your information* $\underset{\textbf{odp:Agent}}{\underline{\textit{with our service providers}}}$
> $\underset{\textbf{odp:Process}}{\underline{\textit{for the purpose of receiving services such as security, fraud detection}}}$
> $\underset{\textbf{odp:Process}}{\underline{\textit{and prevention, reporting, and storage.”}}}$

the whole underlined proposition, associated with the *odp:Process* concept, is detected from the noun chunks *purpose* and *security* extracted by the vocabulary-driven approach and mapped on the *Purpose* and *Security* DPV concepts, respectively.

Moreover, in those cases where the vocabulary-driven concepts extraction does not detect any purpose of data processing, a mention to this information is looked in the optional constituents of a clause. In this case, the system looks for an adverbial that is introduced by the preposition *for*, that was found to co-occur with several verbs evoking processing activities, as shown in Table 4.16. The analysis of the co-occurrences of verbs and optional constituents also highlighted a slightly different behaviour of the verb *use*. Indeed, when it is found in a SVO clause, many of its optional constituents are labelled with the XCOMP dependency relation, that represents a predicative complement without its own subject. To take into account this finding, when the system detects a clause with the verb *use*, it first looks for the presence of an adverbial introduced by the particle *for*. If this constituent is not present, then it looks for a XCOMP constituent for representing the purpose of the processing. For instance, from the following sentence:

> *“We* $\underset{\textbf{dpv:Use}}{\underline{\textit{use}}}$ *your personal information*
> $\underset{\textbf{odp:Processing}}{\underline{\textit{to recommend features, products and services}}}$
> *that might be of interest to you”.*

the vocabulary-driven approach fails to extract the processing purpose, because it is not expressed by a noun phrase. However, the ClausIE tool extracts the SVO clause *(S: we, V: use, O:information, XCOMP: recommend)*, where the presence of the XCOM constituent allows the detection of the verbal proposition that describes the purpose of the processing.

The output of the vocabulary-driven extraction is also used to find mentions to personal data within the sentence that describes a data processing scenario. In this case, the intuition was to exploit the mappings between noun chunks and concepts in the DPV *Personal Data Category* module. Indeed, the information modelled by that module corresponds to the information modelled by the *Personal Data* class in the *Privacy Policy Personal Data* pattern. Specifically, the system checks if three or more noun chunks are associate with some concepts in the *Personal Data Category* module of the DPV. If this is the case, then it is assumed that the presence of multiple noun chunks mapped on DPV concepts supports the assumption of a correct extraction performed by the vocabulary-driven approach. By contrast, if the number of personal data extracted from a sentence is less than three, then the system relies on the synthetic clauses for determining whether the sentence excerpts correspond to personal data. The synthetic clauses the system looks for are made of a subject *you*, an inferred verb *“has”* and an object that corresponds to the sentence excerpt that was found by the vocabulary-driven extraction. The implementation of this control is based on the results obtained in the OIE task, where synthetic clauses with those characteristics were found to frequently occur in the sentences, as was also highlighted in Table 4.3. Considering, again, the sample sentence:

> *“We* $\underset{\textbf{dpv:Use}}{\underline{\textit{use}}}$ $\underset{\textbf{odp:PersonalData}}{\underline{\textit{your personal information}}}$ $\underset{\textbf{odp:Processing}}{\underline{\textit{to recommend features}}}$ $\underset{\textbf{odp:Processing}}{\underline{\textit{products, and services}}}$ *that might be of interest to you.”*

ClausIE extracts from it the synthetic clause *(you, “has”, information)* that, combined with the output of the vocabulary-driven approach which extracted the *personal information* chunk, contributes in identifying a reference to personal data in the text.

Table 4.18 summarises the information about the clauses and the DPV mappings that are considered for detecting the concepts in the *Privacy Policy Personal Data* pattern. The final result of this approach is a set of sentences that describe a data processing scenario characterised by several information. In those sentences, distinct text excerpts were detected by leveraging the information modelled in different ontological resources and some regularities that were found by analysing their syntactic structure.

In this analysis, the concepts *Lawful Basis For Processing*, *Data Collection Step* and *Data Sharing Step* and *Duration* were not included. As already mentioned, the *Lawful Basis For Processing* were not taken into consideration because the OIE task and the vocabulary-driven extraction were tested on the OPP-115 corpus, where those concepts are not present due to the US privacy

DPV concept in the SVO clause	odp:Agent	odp:Collection Mechanism	odp:Processing	odp:Personal Data
Obtain - Acquire, Collect, Record	A: from	A: automatically	mappings on DPV A: for	mappings on DPV synthetic clauses
Disclose - DiscloseByTransmission, Disseminate, MakeAvailable, Share, Transmit	A: with A: to	*not applicable*	mappings on DPV A: for	mappings on DPV synthetic clauses
Store	*not applicable*	*not applicable*	mappings on DPV A: for	mappings on DPV synthetic clauses
Use - Analyse, Consult, Profiling, Retrieve	*not applicable*	*not applicable*	mappings on DPV A: for XCOM constituents	mappings on DPV synthetic clauses

Table 4.18: The information from the clauses and the mappings with the DPV that were used to detect specific classes (preceded by the odp prefix) in the *Privacy Policy Personal Data pattern* withing the sentences.

legal framework of reference. By contrast, no particular lexico-syntactic evidence about the other mentioned concepts emerged from the analysis of the clauses. This suggests that more sophisticated NLP techniques should be applied for detecting this information in privacy policies. This need has been acknowledged and inserted as a future improvement, as it will be discussed in Chapter 7. The next chapter, instead, will present the results obtained from the execution of the described approach on a corpus of privacy policies on which the legal framework set by the GDPR applies. Moreover, it will present the manual annotation task that was implemented for evaluating the results. Finally, a discussion about the findings, the limitations and possible future improvements of the approach will be presented.

4.3 Summary

This chapter described the experimental part of the thesis, that led to the implementation of a system for detecting recurrent information scenarios from the text of privacy policies. Those information scenarios refer to data processing activities identified based on the concepts and the relations modelled in a domain specific ODP, i.e. the *Privacy Policy Personal Data* pattern. To identify those scenarios, the system adopts different NLP approaches and the information in domain-specific and domain-independent vocabularies. On the one hand, with an open information extraction approach, it detects recurrent lexico-syntactic patterns in the sentences. On the other hand, with a vocabulary-driven approach based on text similarity, it extracts mentions to domain-specific concepts in the sentences. Combining those approaches, the system extracts those sentences that describe data processing scenarios characterised with respect to the concepts modelled in the *Privacy Policy Personal Data* pattern.

5 Evaluation and Results Discussion

This chapter presents the evaluation of the integrated approach for the extraction of processing scenarios from the text of privacy policies. The system has been tested on a GDPR-oriented set of privacy policies, selected from a recently-released corpus, named Princeton-Leuven Longitudinal corpus.

The first part of the chapter describes the criteria that were used to select from that corpus the 25 privacy policies on which the implemented system has been applied to extract recurrent processing scenarios.

The second part of the chapter explains how the annotation task was designed in order to gather the experts' assessment about the output of the system. Finally, the performance of the system is presented and discussed with respect to the assessment provided by the legal experts.

5.1 Test Set Construction

5.1.1 The Princeton-Leuven Longitudinal Corpus

The Princeton-Leuven Longitudinal corpus[1] [6], released on March 2020, is a dataset of English privacy policies that were issued in the last two decades. The documents in the corpus were collected automatically through a Web crawler that searches for privacy policies in the Web pages stored in the Internet Archive's Wayback Machine.

Overall, the dataset is made of 910.546 privacy policies from 108.499 Web sites that were included in the top-100K Alexa rank at least once in the years between 2009 and 2019. For each Web site, two versions per year of its privacy policy are available: one released in the first half of the year and one released in the second half. Each privacy policy is associated with one or more category labels, chosen from a set of 16 possible labels that represent the offered service or the main activity of the Web site to which the privacy policy applies.The privacy policies are available both in Markdown formatted text and HTML format,

[1] https://privacypolicies.cs.princeton.edu/

Category	Web Site
Business	salesforce.com (105, USA), yelp.com (232, USA), zendesk.com (248, USA), fiverr.com (426, ISR), hootsuite.com (540, CAN)
Education	sciencedirect.com (322, GBR), coursera.com (416, USA), mit.edu (432, USA), livescience.com (2889, USA), macmillandictionary.com (2918, GBR)
Information Tech	thedoctopdf.com (4338, ISR), developer.wordpress.com (5058, USA), ablebits.com (5190, BLR), anandtech.com (6531, USA), picresize.com (7850, USA)
Entertainment	fandom.com (84, USA), mydramalist.com (1613, not found), musixmatch.com (2640, ITA), itv.com (2729, GBR), digitalspy.com (3074, GBR)
Shopping	amazon.co.uk (82, USA), gearbest.com (320, CHN), zaful.com (1530, HKG), redbubble.com (1704, USA and AUS), rei.com (2820, USA)

Table 5.1: The 25 policies from the Princeton-Leuven corpus that were selected as a test set. Policies are grouped by category label. In the second column, the address of a Web site is associated, in brackets, with its Alexa rank and the code of the country where the company to which the site pertains is established.

cleared of unnecessary tags and text from headings, footers and sidebars in the Web page.

5.1.2 Privacy Policies Collection and Processing

The Princeton-Leuven Longitudinal Corpus was the starting point for the collection of a set of privacy policies on which the legal framework set by the GDPR applies. The first decision regarding the characteristics of the dataset to be collected has concerned the number of documents to be included. Because I planned a manual assessment of the results, made by two legal experts, the size of the dataset should have been such to not make the task too burdensome for the experts. For this reason, I chose to include 25 documents in the dataset, consistently with other state-of-the-art approaches that were executed or validated on GDPR-oriented corpora on which some task of manual expert annotation was executed[2].

After deciding on the size of the dataset, I considered the statistics about the category labels distribution provided by the Princeton-Leuven corpus, showing that most of the privacy policies are associated with few dominant labels in the dataset [6]. Based on those statistics, I decided to select five privacy policies for each of the five most frequent categories. First, to choose the privacy policies, I considered the top-150 Web sites for each category, according to the Alexa's ranking for the second half of 2019. For those Web sites, I retrieved

[2]In PrivacyGuide a set of 45 privacy policies was annotated manually by legal experts to train some classification models, while in KnIGHT the performance of the system is assessed manually on a corpus of 20 privacy policies. The systems are explained in detail in the next chapter, in Section 6.4.5.2 and in Section 6.4.4.1 respectively.

their privacy policies verifying that the time-stamp in the corpus referred to the same time-frame. The setting of these parameters ensured the retrieval of those privacy policies that were released after the entry into force of the GDPR. Second, I analysed manually the selected documents to verify to which of them the GDPR actually applies. In each privacy policy, based on the territorial scope of the Regulation (Art. 3), I looked for an explicit reference to the GDPR or a mention to individuals in the EU as addressees of the good/service offered by the company, regardless of its country of establishment. I read the privacy policies by increasing ranking value and I put in the final set of GDPR-oriented privacy policies the first five documents, in each category, that satisfied one of the aforementioned conditions. The list of selected privacy policies, grouped by category, is shown in Table 5.1.

Once the set of privacy policies was collected, the documents went through a pre-processing step. Starting from the Markdown formatted text, each paragraph of a privacy policy was automatically split into individual sentences, using the Spacy Python package. To handle the bulleted lists inside the text, I followed the approach used by Harkous et al. in their Polisis framework[3] [93], where the processing of bulleted lists varies according to the length of the textual fragments in each item of the lists. Coherently with their approach, the list items having up to 20 words were combined with the introductory statement of the list, based on the assumption that short list items typically are not self-contained. List items having more than 20 words, instead, were treated as self-contained statements, independent form the introductory sentence of the list.

5.1.3 Statistics about the Test Set

The twenty-five privacy policies in the dataset are made of 3691 sentences overall, yielding the average number of sentences per privacy policy to 148 (omitting from the counts the paragraph headings). The privacy policies belonging to the *business* category are, on average, those with the highest number of sentences, being equal to 165. By contrast, the privacy policies in the *shopping* category are the shortest, with 119 sentences on average. The longest privacy policy in the corpus is from the *developer.wordpress.com* Web site, in the *information tech* category, containing 260 sentences, whereas the shortest is from the *mit.edu* Web site, in the *education* category, containing only 46 sentences.

The sentences in the corpus contain 2290 distinct content words (unigrams), increasing to 41981 considering repetitions. The tag cloud in Figure

[3]The Polisis framework is discussed with more detail in the next chapter, in Section 6.4.1.1.

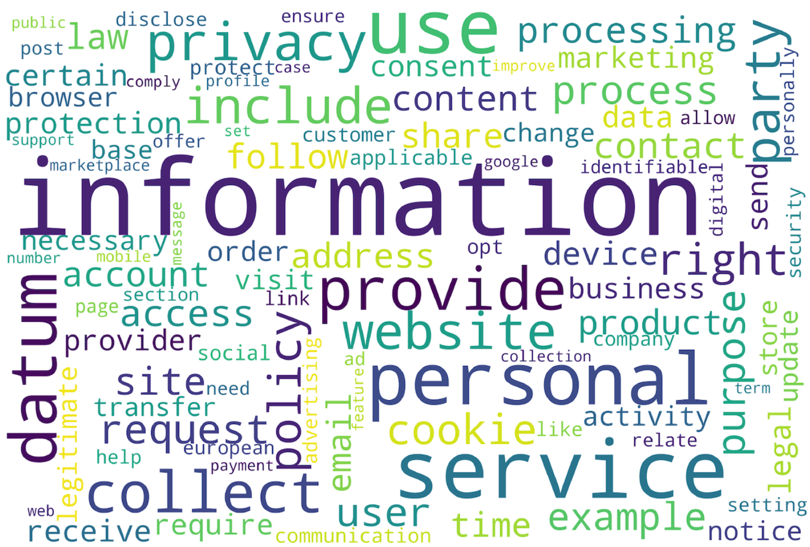


Figure 5.1: The tag cloud with the 100 most frequent content words in the GDPR-oriented dataset of privacy policies.

5.1 shows the 100 most frequent content words across the corpus. Among those words, it can be noticed the presence of the verbs of the DPV *Processing* module that were used by the integrated approach to find sentences describing a processing scenario. The term *use* occurs 1092 times in the corpus[4], being the third most frequent word, after *information* and *service* which occur 2388 and 1246 times, respectively. The verb *collect* is the seventh most frequent term, occurring 617 times in the corpus. With lowest frequencies, but still in the top-100, also the verbs *share*, *store*, *disclose*, *obtain* are included in the tag cloud. In it, there are also some words that were extracted frequently from the vocabulary-driven concepts extraction approach (see Table 4.15 in Chapter 4), applied to the privacy policies of the OPP-115 corpus. Among them, *email* and *contact* are likely to refer, in the privacy policies, to categories of personal data being processed, while *purpose* and *security* are likely to refer to purposes of the processing.

This preliminary analysis of the terms occurring in the GDPR-oriented test set revealed potential correspondences with some of the DPV concepts utilised to guide the OIE task and the text chunks found by the vocabulary-driven concepts extraction applied to the OPP-115 corpus. The next section will present the statistics about the extractions obtained from the integrated approach to

[4]It could be noticed, however, that the term *use* is not disambiguated in the tag cloud. Consequently, it could be either a noun or a verb.

the detection of processing scenarios, providing further insights about the words in the test set.

5.2 Detection of Data Processing Scenarios from the Test Set

The integrated approach to the detection of processing scenarios, described in Section 4.2.5 of Chapter 4, was applied to the GDPR-oriented corpus. The execution of the first step of the method resulted in the detection of 380 sentences that could potentially describe a data processing scenario. The selected sentences are those made of a SVO clause in which the subject is the pronoun *we* and the verb matches the name of a class in the *Processing* module of the DPV. Among them, 171 verbs match the name of the *Obtain* class and 88 verbs match the name of the *Disclose* class, or one of their subclasses. The verb *use*, that matches the homonym class in the DPV, was found 162 times in the SVO clauses of the sentences. By contrast, no term matching the name of a subclass of the *Use* concept was found in the corpus. Finally, the verb *store*, matching the homonym class in the DPV was found in 20 sentences. The third column of Table 5.2 reports the details about the verbs that were extracted from the clauses and utilised by the system to select the sentences referring to a data processing scenario.

It should be noticed that some of the selected sentences contain more than one SVO clause in which the verb matches the name of a class in the DPV. It is the case, for instance, of the following sentence:

> *"We store and use the information you provide during that process, such as the first and last name you enter."*

from which the system detects the SVO clauses *(we, store, information)* and *(we, use, information)*, considering them as evidence of two distinct processing scenarios. This also explains why the number of matching verbs, equal to 441, found in as many SVO clauses exceeds the number of selected sentences.

The second step of the process jointly analysed the optional constituents of the SVO clauses and the output of the vocabulary-driven concepts extraction in order to find text excerpts characterising the processing scenarios described by the selected sentences. The execution of this step revealed that, among the 380 selected sentences, only for 225 of them the system identified at least one text excerpt that allows a more detailed characterisation of the processing scenario described in them. A manual analysis of those sentences showed that they usually refer to introductory statements, like the following one:

DPV Superclass	Verb in the SVO clauses	Frequency in clause (after execution of the first step)	Frequency in clause (after execution of the second step)
Obtain	collect	158	98
	obtain	13	10
Disclose	share	57	44
	disclose	29	18
	transmit	2	*not found*
Store	store	20	7
Use	use	162	83

Table 5.2: The classes in the DPV *Processing* module and the verbs in SVO clauses that lexically matched the names of those classes (or their subclasses).

	Frequency	Most frequent sentence excerpt	Most frequent mapping with DPV
odp:Agent (collection scenario)	47	"from you" (16)	*not applicable*
odp:Agent (disclosure scenario)	36	"with third parties" (5)	*not applicable*
odp:CollectionMachanism	9	"automatically" (9)	*not applicable*
odp:Processing (with DPV mappings)	183	"for a variety of purpose" (21)	dpv:Purpose (105)
odp:Processing (without DPV mappings)	13	"to provide you with the product key and support service that you order from us" (1)	*not applicable*
odp:PersonalData	199	"personal information" (23)	dpv:DeviceBased (42)

Table 5.3: Statistics about the mappings found by the integrated approach to the detection of processing scenarios' details.

> *"This privacy notice explains how we collect and use your personal information if you are a customer of our products and services, for example if you use our websites."*

The impossibility for the system to detect any detail of a processing scenario can be thus seen as an evidence of a generic sentence which does not describe a meaningful scenario with respect to the information modelled in the *Privacy Policy Personal Data* pattern. Consequently, those sentences were discarded from further the analysis.

In the remaining 225 sentences, 199 sentence excerpts were identified as referring to a personal data. Those sentence excerpts were mapped by the vocabulary-driven approach on 37 concepts in DPV *Personal Data Category* module. Furthermore, 196 sentence excerpts were identified as referring to the *purpose* of a data processing. Among them, 183 sentence excerpts were mapped on 11 classes of the *Purpose* module of the DPV, while 13 sentence excerpts were detected by the analysis of the optional constituents of the clauses, consequently, they have no mapping with the classes in the DPV.

The system also detected 83 sentence excerpts referring to an agent involved in a data processing activity and 9 sentence excerpts containing an explicit mention to the automatic mechanism for collecting data. Table 5.3 sum-

DPV concept in the SVO clause	odp:Agent	odp:Collection Mechanism	odp:Processing	odp:PersonalData
Obtain - dpv:Collect	A:from	*not applicable*	mappings on DPV	*not found*

The information we [collect]**dpv:Collect** [from you]**odp:Agent/A:from** may be used in the following ways: [to improve customer service]**odp:Processing/dpv:CustomerCare** (your information helps us to more effectively respond to your customer service requests).

DPV concept in the SVO clause	odp:Agent	odp:Collection Mechanism	odp:Processing	odp:PersonalData
Disclose - dpv:Share	A:with	*not applicable*	*not found*	mappings on DPV

We may [share]**dpv:Share** information from or about you (such as your city, and if you provide it, your [age]**odp:PersonalData/dpv:Age** and [gender] **odp:PersonalData/dpv:Gender**), your [device type]**odp:PersonalData/dpv:DeviceBased**, and your use of the Service (such as which businesses you bookmark or call, or if you visit a business URL) [with businesses] **odp:Agent/A:with** on Yelp.

DPV concept in the SVO clause	odp:Agent	odp:Collection Mechanism	odp:Processing	odp:PersonalData
Store - dpv:Store	*not applicable*	*not applicable*	*not found*	mappings on DPV

We may also [store]**dpv:Store** your [location]**odp:PersonalData/dpv:Location** whenever our mobile applications are running, including when running in the background, if you enable our mobile apps to access such information in the course of using the Service.

DPV concept in the SVO clause	odp:Agent	odp:Collection Mechanism	odp:Processing	odp:PersonalData
Use - dpv:Use	*not applicable*	*not applicable*	XCOM constituent	*not found*

We [use]**dpv:Use** information about you [to help us understand usage patterns and other activities on our websites and applications so that we can diagnose problems and make improvements, including enhancing usability and security.] **odp:Purpose/ XCOM_constituent**

Figure 5.2: Examples of the sentences extracted with the implemented approach. The headings of the four tables in this figure correspond to the heading of Table 4.18 in Chapter 4. Each table in this picture provides a reference to the specific lexical-syntactic evidence, among those in Table 4.18, that was used to map a sentence excerpt on the corresponding class of the *Privacy Policy Personal Data* ODP. In the sentences below the tables, the text chunks that were mapped on the classes of the ODP are enclosed between brackets. Each text chunk is also associated with the information about the ODP class on which it is mapped and the lexico-syntactic evidence that generated the mapping. The verbs that were used to select the SVO clauses in the first step of the implemented approach (see Section 4.2.5.2) are only associated with the information about the corresponding concept in the DPV *Processing*.

marises those data, indicating the classes of the *Privacy Policy Personal Data* pattern to which the sentence excerpts correspond and providing some examples. From the last column of the table, it can be noticed that the majority of the sentence excerpts that refer to the purpose of a data processing activity, where mapped on the *Purpose* class, that is the most general class that the DPV uses to represent the purposes of the processing. Those general mappings can be explained considering the weakness that was identified in the vocabulary-driven approach in identifying purposes of the processing that are described with phrases that are more articulated than simple noun chunks. By contrast, the mappings between sentence excerpts and concepts in the *Personal Data Category* module are scattered on several classes, consistently with the intuition that personal data types are usually described with short noun phrases that can be more easily recognised by the vocabulary-driven approach.

Figure 5.2 provides some examples of the sentences that were extracted with the implemented approach and the corresponding mappings established with the classes of the *Privacy Policy Personal Data* pattern.

The next sections will describe how the performance of the system was evaluated by two legal experts through a manual assessment of the processing scenarios extracted by the system.

5.3 Evaluation by Legal Experts

5.3.1 Objective of the Evaluation Task

As mentioned earlier in the chapter, the evaluation of the results was assigned to two legal experts, both having a Master degree in Law, who were asked to express a judgement about the characterisation of some processing scenarios extracted from the GDPR-oriented corpus.

The definition of the manual annotation task involved the identification of the evaluation objectives, determining which aspects of the output produced by the implemented pipeline should have been evaluated. Indeed, the process that leads to the characterisation of the processing scenarios described in privacy policies is the result of several conceptual and implementational choices, motivated by the intent to re-use the ontological model provided by the *Privacy Policy Personal Data* pattern and the taxonomies of concepts in the DPV modules.

Considering those conceptual and modelling choices, I identified three evaluation objectives:

(o1) assessment of the precision of the method in selecting the sentences of a privacy policy that describe data processing scenarios. This selection is based on the lexical correspondence between a verb in a SVO clause and the name of a class in the DPV *Processing* module;

(o2) assessment of the precision of the method in characterising a data processing scenario, based on groups of concepts that were identified in the *Privacy Policy Personal Data* pattern and some correspondences made with processing activities modelled in the DPV *Processing* module (recall Table 4.16 in Chapter 4). At the implementation level, this step relies on the analysis of both the essential and the optional constituents of the clauses, integrating the information about sentence excerpts extracted by the vocabulary-driven approach;

(o3) assessment of the precision of the mappings between sentence excerpts and concepts in the DPV, considering that the integrated approach to the detection of processing scenarios was designed to filter out inaccurate mappings extracted by the vocabulary-driven approach.

The next section will explain how the annotation task has been designed, for allowing the assessment of the system with respect to the proposed objectives.

5.3.2 The Annotation Task

5.3.2.1 Design of the Annotation Task with respect to the Evaluation Objectives

The task was explained to the annotators introducing the concept of *processing template* and *template components*. A processing template corresponds generically to a data processing scenario, whereas the *template elements* characterise the scenario with respect to the information modelled in the *Privacy Policy Personal Data* pattern.

The annotators were provided with the description of three processing templates: *obtain*, *disclose*, *other processing*. Given a sentence under evaluation, the *obtain* template was associated with that sentence when the system detected a SVO clause whose verb lexically matched the name of a class, or one of its subclasses, in the DPV *Processing* module. Similarly, the *disclose* template was associated to a sentence made of a SVO clause in which the verb lexically matched the *Disclose* class, or one of its subclasses. The *other processing*, instead, was associated to a sentence under evaluation when the system detected a SVO clause in which the verb lexically matched the class *Store* or *Use*.

The template components associated with each processing template correspond to the details about a processing scenario that the system tries to identify in the second step of the integrated approach, as explained in the previous chapter (see Section 4.2.5.3). A further component, named *processing type*, was added to each template. This component indicates to the experts the verb that yielded the association of the sentence under evaluation with the corresponding processing template. Table 5.4 summarises the names of processing templates and their components, showing the correspondences with the classes in the *Privacy Policy Personal Data* pattern.

Based on the identified evaluation objectives and having provided the experts with the explanations related to the processing templates, the annotation task was designed as a three-levels questionnaire. In each level, the expert is asked to answer a question about the processing template associated with a sentence to be evaluated. There are only two possible answers to the questions:

	Obtain	Disclose	Other processing
Processing Type (dpv: Obtain, dpv:Disclose, dpv:Store, dpv:Use, including the subclasses)	X	X	X
Personal data (odp:PersonalData)	X	X	X
Purpose (odp:Processing)	X	X	X
Obtains data from agent (odp:Agent)	X		
Discloses data to agent (odp:Agent)		X	
Mechanism for obtaining data (odp:CollectionMechanism)	X		

Table 5.4: The processing templates with the corresponding template components that were proposed to the experts. The first row shows the names of the processing templates, whereas the names of the template components are listed in the first column. The name of each template component is associated with the corresponding class, in the DPV or in the *Privacy Policy Personal Data* pattern, that justifies the presence of the template component.

yes or *no*. The questions were formulated to focus on a gradually increasing level of detail across the three levels.

The question at the first level in the task is the most generic one. Given the sentence under evaluation and given the label of a processing template, the question put to the experts is formulated as follows:

Is the processing template appropriate to represent the information expressed in the sentence? (Q1)

This question had a twofold objective. First, to familiarise the expert with the sentence to be evaluated, by inducing her to read the whole sentence at least once, before moving on to the other ones, referring to individual sentence excerpts. Second, to allow the assessment of the system performance with respect to the first evaluation objective (identified as **(o1)** in the previous section), i.e. the system precision in identifying sentences that describe a data processing scenario.

Keeping the focus on the current sentence under evaluation, the questions at the second and the third level in the annotation task focused on the assessment on the distinct templates elements for which the system found a corresponding sentence excerpt. Specifically, the second question is formulated as follows:

Does the sentence excerpt express the information represented by the template component? (Q2)

This question aimed to grasp the experts' assessment about the second evaluation objective (**o2**), i.e. the system performance in characterising a processing scenario.

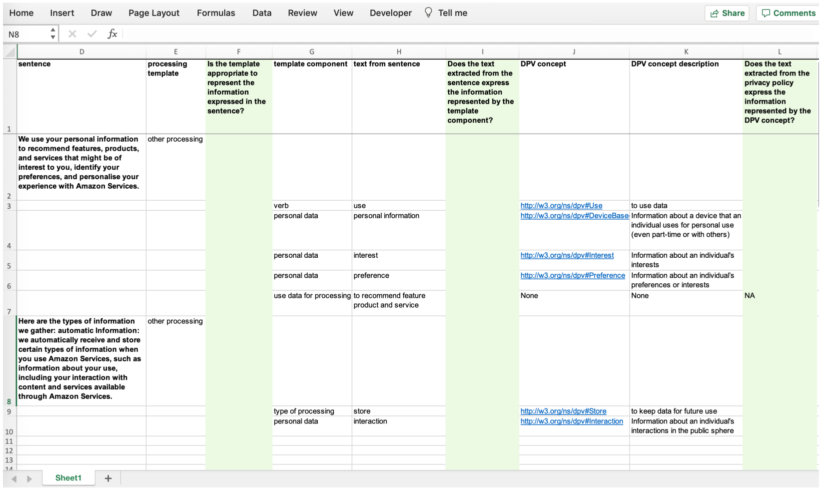
Referring to the same template component and the third evaluation objective **(o3)**, the third and last question, focused on the mappings between the text excerpt associated to the template component under evaluation and the concept in the DPV. Specifically, given the sentence excerpt and given the DPV concept with its description, provided in the vocabulary itself, the question was formulated as follows:

Does the sentence excerpt express the information represented by the DPV concept? (Q3)

This question as been considered applicable only to the sentence excerpts where the mappings with the DPV concepts were available and were extracted by the vocabulary driven approach. The second and the third questions were repeated for each component of the template associated to a sentence, before moving to the next sentence.

In each level of the annotation, if the expert provided a negative answer to the question, then she was instructed to skip all the questions at the following levels, for the sentence under evaluation. Consequently, a negative answer in the first level of the annotation, invalidated all the answers for the questions at second and third level. Hypothetically, the three questions could have been answered with a certain degree of independence. Indeed, even if the association of a sentence with a template were to be considered incorrect, the expert could express a judgement regarding the components that are common to all processing templates (i.e. *processing type*, *personal data*, and *purpose*). Similarly, she could express a judgement about the mapping between sentence excerpts and concepts in the DPV. However, I have chosen the option that invalidates all steps following a question with negative answer, in order to make the explanation of the annotation task to the experts simpler and more intuitive.

The experts were also instructed, in question Q2, to consider valid sentence excerpts that only provide a shallow mention to the information represented by the template component they are associated with. This is the case, for instance, of a sentence starting with *"We collect your personal data..."*, in which the sentence excerpt *personal data* could be associated with the homonym template component *personal data*. The association could be deemed valid considering the objective of evaluating the performance of the system in correctly associating sentence excerpts with the corresponding template components. Indeed, I assumed that, if the system can correctly identify the text chunk *personal data*, then it could also detect more specific mentions to personal data types in a sen-



sentence	processing template	Is the template appropriate to represent the information expressed in the sentence?	template component	text from sentence	Does the text extracted from the sentence express the information represented by the template component?	DPV concept	DPV concept description	Does the text extracted from the privacy policy express the information represented by the DPV concept?
We use your personal information to recommend features, products, and services that might be of interest to you, identify your preferences, and personalise your experience with Amazon Services.	other processing							
			verb	use		http://w3.org/ns/dpv#Use	to use data	
			personal data	personal information		http://w3.org/ns/dpv#DeviceBase	Information about a device that an individual uses for personal use (even part-time or with others)	
			personal data	interest		http://w3.org/ns/dpv#Interest	Information about an individual's interests	
			personal data	preference		http://w3.org/ns/dpv#Preference	Information about an individual's preferences or interests	
			use data for processing	to recommend feature product and service		None	None	NA
Here are the types of information we gather: automatic Information: we automatically receive and store certain types of information when you use Amazon Services, such as information about your use, including your interaction with content and services available through Amazon Services.	other processing							
			type of processing	store		http://w3.org/ns/dpv#Store	to keep data for future use	
			personal data	interaction		http://w3.org/ns/dpv#Interaction	Information about an individual's interactions in the public sphere	

Figure 5.3: The spreadsheet that was provided to the annotators. The expert was asked to insert her annotation in the columns highlighted in green. Some rows in rightmost green column were filled with a pre-defined NA (i.e. not applicable) value when the no mappings with the DPV concepts were found or when the sentence was not applicable to the template element.

tence having the same lexico-syntactic structure, e.g. *"we collect your contact information"*. For the same reason, the experts were instructed to evaluate sentences regardless from the fact that the described processing activity is stated as being performed or not. This could be the case, for instance, of a negative sentence starting with "We do not collect your personal information" . The next section explains how the sentences to be assessed by the annotators were chosen.

5.3.2.2 Set up of the Documents to Provide for Performing the Annotation

To collect the answers to the questions, I prepared a spreadsheet to provide to the experts, organising it so that they were guided through the various levels of the annotation. The look of the spreadsheet is shown in Figure 5.3.

Information about the annotation task and the instructions to fill the spreadsheet were written in a document containing the guidelines to be followed. The full version of the document is available in Appendix B, whereas this section briefly summarises its main structure.

The document is divided into two sections. The first section recalls some of the main concepts of the GDPR that are relevant with respect to the annotation task. Moreover, it describes the DPV vocabulary in a technology-neutral language, to make comprehensible the main aspects of the vocabulary without asking for expertise in the fields of knowledge representation and Semantic

Web. The description also relies on graphical representations of the vocabulary modules concerned by the annotation, to ease understanding of the role of the resource in the annotation task.

The second section of the guidelines document explains the annotation task. The reader is introduced to the concepts of type template and template components and she is provided with their descriptions. Next, she is guided through the three-level annotation. For each level, the guidelines document specifies the question to be answered and the admissible answers for the question. Furthermore, the explanation is complemented with a screenshot of the spreadsheet that highlights the cell in which the answer is to be typed. Moreover, a draft version of the document that I prepared, was also provided to the experts, with the annotation of two sample sentences.

The guideline document was also used to estimate the amount of time required for the annotation and assess the number of sentences to be assigned to the experts for the evaluation. Details about the criteria used to estimate that number are provided in the next section.

5.3.2.3 Selection of the Sentences to be Annotated

The evaluation task was performed by two experts on a voluntary basis and they did not receive any remuneration for the time spent in performing the task. In order not to make the work too heavy, I chose not to ask the experts to annotate all the templates extracted from the 225 sentences selected by the system, but only some of them. In particular, I envisaged a maximum effort to complete the task equal to 8 hours (corresponding to one working day in Italy).

Before assigning the task to the experts, I tried to estimate the time for reading the guidelines document and for performing the annotation task. The times were estimated as follows:

- 1 hour and a half for reading and understanding the annotation guidelines;
- 7 minutes for each of the first 10 sentences under evaluation in the spreadsheet, for providing the three-level annotation. It was assumed that, at the beginning of the task the expert needs to familiarise with her job, needing further readings of the information provided in the guidelines document,
- 6 minutes for answering the questions for each of the 10 following sentences under evaluation in the spreadsheet, assuming an increased confidence of the expert in the performance of the task.

According to those estimations, the effort for reading the guidelines and annotating 20 sentences would have take just over 3 hours and a half. At that point, assuming a consolidated confidence and familiarity of the expert with the task, I assumed that the time required for answering the questions of a sentence under evaluation could have been further reduced to about 5 minutes per sentence. According to those estimates, in the remaining four and a half hours, I considered feasible the annotation of another 55 sentences for the experts (with an average time of 4.7 minutes for sentence).

The experts were asked to track the actual time that took to accomplish each group of sentences, to evaluate the differences between the estimated and the actual times for annotation.

To choose the sentences to be annotated, I decided to look for the sentences containing a minimum set of details about the processing scenario. Specifically, I choose those criteria:

- for the sentences referring to the *obtain* template, the system detected at least one personal data or one purpose, together with the mention to the agent that provides that personal data or the collection mechanism;
- for the sentences referring to the *disclose* template, the system detected at least one personal data or one purpose, together with the agent to which the personal data are disclosed;
- for the *other processing* template, the system detected at least a personal data or one purpose of the processing.

Based on those criteria, 128 sentences were selected, among which 27 associated to the *obtain* template, 11 associated to the *disclose* template and 90 associated to the *other processing* template. Thus, to balance the number of templates, I inserted in the set of sentences to be annotated all those associated with the *obtain* and the *disclose* templates. For the *other processing* template, I tried to insert in the dataset at least one sentence for each privacy policy, randomly selecting, with the Python `random` package, additional sentences from those privacy policies in which more that two sentences related to the *other processing* template were extracted.

5.3.3 Assessment of the Reliability of the Performed Annotation Task

Before looking at the results obtained from the annotation task, some preliminary remarks can be drawn about the timing indicated by the experts for completing the task and the difference from the estimates that I made. The results are summarised in Table 5.5.

	Initial Estimate	$Expert_1$	$Expert_2$
Time for reading the guidelines	90.0 min	120.0 min	60.0 min
Time for annotating one of the first 10 templates	7.0 min	4.0 min	5.5 min
Time for annotating one of the following 10 templates	6.0 min	2.5 min	4.0 min
Time for annotating one of the remaining 55 templates	4.7 min	1.5 min	2.0 min
Total time	479.0 min	267.5 min	265.0 min

Table 5.5: Statistics about timings in reading the guidelines and performing the annotation task

According to the provided data, the time indicated by the first expert to read the annotation guidelines is twice the time indicated by the second expert. However, the average time it took the two experts to read the guidelines is consistent with the initial estimate that I made about reading the document. By contrast, the time for performing the annotation of the 75 processing templates was less than the estimates that I initially assumed. In this case, the first expert reported annotation times that are lower than those reported by the second expert, for all three sets of annotated templates. However, summing the times for reading the guidelines and for annotating the templates, the resulting time effort is similar for both the experts and it is approximately equal to four hours and a half. The estimates I have made differ from the actual times required to perform the annotation especially when considering the time to annotate the last 55 templates. In this case, the data provided by both experts show that, after annotating the first 20 processing templates, the familiarity with the task is consolidated and the annotation task speeds up.

Details about the answers provided by the experts in the annotation task are summarised in Table 5.6 and discussed hereafter. In the table, the statistics about the 75 sentences under evaluation provided by the annotators are shown in the second column. As already mentioned, the test set contained 75 processing templates, for which the experts provided as many assessment by answering question Q1 (see Section 5.3.2.1). Overall, the templates are made of 235 templates components, that were assessed by the experts by answering question Q2. Among the sentence excerpts that were associated to the *processing type*, *personal data* and *purpose* template components, 192 of them where associated with a concept in the DPV *Processing*, *Personal Data Category* and *Purpose* modules, respectively. For those mappings, the experts expressed their evaluation by answering the question Q3.

When calculating the agreement between annotators and the performance of the system, I had to take into account the guidelines that the experts were provided with, particularly referring to the instruction of interrupting the annotation of a template when they provided a negative answer to one of the three questions. Consequently, with the exception of the first question, where

	Overall number in the test set	Overall number of answers considered given the answers to the previous step	Num. answers agreeing (both "yes" and "no")	Cohen Agreement	Num. positive answers agreeing (only "yes")	Overall precision
Q1	**75**	**75**	**66**	**0.41**	**62 (88.0%)**	**88%**
obtain template	27	27	26		26	
disclose template	11	11	11		11	
other processing template	37	37	29		25	
Q2	**235**	**205**	**196**	**0.84**	**168 (82%)**	**71.5%**
processing type	75	62	62		62	
personal data (odp:PersonalData)	57	56	48		37	
purpose (odp:Processing)	65	50	49		33	
obtains data from agent (odp:Agent)	19	18	18		17	
discloses data to agent (odp:Agent)	11	11	11		11	
mechanism for obtaining data (odp:CollectionMechanism)	8	8	8		8	
Q3	**192**	**131**	**111**	**0.45**	**100 (73.3%)**	**52.1%**
processing type (dpv:Processing)	75	62	62		62	
personal data (dpv:PersonalDataCategory)	57	37	35		24	
purpse (dpv:Purpose)	60	32	14		14	

Table 5.6: The results of the evaluation made by the experts. The first and the second columns represent the elements to be evaluated in each level of the annotation, with the corresponding number of occurrences in the tests set. The third column reports the number of the overall elements to be evaluated in each step, considering that a negative answer in the preceding step invalidated all the answers in the following steps. The forth column indicated the number of answers in which the experts evaluations agree and the fifth column reports the Cohen's κ coefficient. The sixth column represent the number of answers were both the experts agreed in providing a positive evaluation, based on which the precision of the system was computed, with respect to the overall number of elements, reported in the second column.

the statistics about the evaluation were calculated on all the 75 provided answers, to calculate statistics about the following answers, I only considered the elements that were positively assessed by both experts in the previous step of the annotation. Thus, in the third column of the table, I recorded the number of overall answers that were considered to compute the inter-annotator agreement and the precision of the system.

According to those considerations, I computed the inter-annotator agreement at all three levels of annotation. The measurement of inter-annotator agreement is fundamental in any manual annotation task, where the expressed judgements are affected by the subjectivity of the single individual that is performing the annotation. As noticed by Di Eugenio [65]:

> *"This raises the question of how to evaluate the 'goodness' of a coding scheme. One way of doing so is to assess its reliability, namely, to as-*

> *sess whether different coders can reach a satisfying level of agreement with each other when they use the coding manual on the same data."*

Therefore, the assumption underpinning the measurement of the inter-annotator agreement is that the data are reliable if two or more annotators agree in associating an item under evaluation with a category label (*yes* or *no* in this annotation task). A good level of agreement among annotators is, thus, the precondition to demonstrate the validity of an annotation scheme [195]. The measure that I used to compute the agreement between annotator was the κ coefficient introduced by Cohen and widely adopted in computational linguistics [41]. The κ coefficient is part of the family of chance-corrected coefficients for measuring agreement. It assumes the independence between the annotations provided by the two coders, like it was the case in this task, and presupposes the presence of a prior distribution, unique to each coder, governing the random assignment of categorical labels to the items under evaluation [9]. The κ value can range in the interval [-1,+1], where 0 represents the amount of agreement that can be expected from random chance, and 1 represents perfect agreement between the annotators [127]. Different interpretations of the values that the coefficient can assume have been proposed. One of the best interpretation was that proposed by Landis and Koch [111], that identified ranges of values in the [-1,+1] interval and the strength of the agreement that values in each interval express. According to their interpretation of κ coefficient, negative values represent poor agreement. Positive values, from 0 to 1 are, instead, divided in five regular intervals with step 0.2, that respectively represent *slight*, *fair*, *moderate*, *substantial* and *almost perfect* agreement.

In this annotation task, the κ values computed at each of the three levels, considering in each level only the items that passed the previous step of the annotation, are in all cases greater that 0.40, as indicated in the fifth column of the table. Specifically, the κ value for question Q1 is equal to 0.41, that expresses a moderate agreement, according to the interpretation provided by Landis and Koch.

The agreement on answers provided for question Q2 was computed considering only the components of the 62 processing templates that were annotated with *yes* by both the experts in question Q1. In this case, the κ coefficient reached a value equal to 0.84, that indicates a high level of agreement between the experts and, based on the assumptions underpinning inter-annotator agreement measures, a high level of reliability of the annotation scheme explained to the experts by the guidelines. Finally, the agreement on answers provided for question Q3 was computed considering only the templates components that were assessed positively in question Q2. The κ value in this case is equal to

0.45, indicating a *moderate* agreement. Because the third level of annotation involved the assessment of a mapping between a text chunk and a concept in the DPV, given its description in natural language provided by the vocabulary, this value of the κ coefficient could also be useful for evaluating the effectiveness of the descriptions associated to the concepts in the DPV in their ability to define a concept.

Overall, I considered the obtained values of the κ coefficient as an indicator of the acceptability of the annotation scheme and the consequent reliability of the evaluation provided by the experts to assess the performance of the integrated approach. Considering the lowest κ value obtained by question Q1, the acceptability of the annotation scheme is also confirmed by another interpretation of the κ value that was proposed by Cicchetti and Sparrow [36] and that indicates as *fair* the κ values above 0.40.

Having established the validity of the annotation scheme, the next section will discuss the performance of the system according to the expert assessment.

5.3.4 Results of the Experts' Evaluation

Having ascertained the validity of the annotation scheme, it is now possible to discuss the performance of the system with respect to the evaluation objectives identified in Section 5.3.1.

The first objective **(o1)** concerns the assessment of the precision of the method in selecting sentences that describe a data processing scenario, relying on the lexical matches between the verbs in SVO clauses and the concepts in the DPV *Processing* module. The experts' evaluation about the performance of the system in this regard was investigated by question Q1. The experts agreed in the association of a sentence to a processing template for 62 sentences out of the 75 that were proposed in the annotation task. Therefore, the resulting precision of the system in this first step is equal to 88%. This result shows that the reliance on lexical matches between the verbs in SVO clauses and the names of classes in the DPV *Processing* module can effectively help in detecting sentences describing some processing scenario.

The second evaluation objective **(o2)** focuses on a more detailed view about the processing template and aims to assess the precision of the method in associating specific text excerpts to the appropriate template components. The experts' evaluation referring to this objective was investigated by question Q2. The performance of the system was measured by considering the 205 components of the processing templates that were positively assessed by both experts in the first level of the annotation. Among them, 168 template components received a positive assessment from the experts, resulting in a precision equal to 82%, calculated with respect to the 205 template components that passed the first step of the evaluation. The precision decreases to 71%, when it is computed by considering the overall set of 235 template components extracted by the system and marking as incorrect both extractions in which the experts' judgement disagrees and those in which the experts' judgement agrees in a negative answer.

From the details provided in Table 5.6 about the evaluations expressed for specific templates components, it emerges that the 62 verbs (associated with the *processing type* component), that contributed to link a sentence with a specific processing template, were all positively assessed by both experts at the second level of annotation. This result demonstrates that the reliance on those verbs for identifying processing scenarios is not only useful from an implementative point of view. Instead, those verbs also contribute to the understanding of a processing activity by the reader.

The experts' assessment in this second level of analysis also shows that the lexical match of a verb in a SVO clause and the name of a class in the DPV *Processing* module is not only helpful in selecting sentences that describe a processing scenario. By contrast, it helps in the extraction of the details that characterise a processing activity. Indeed, the lexico-syntactic patterns tailored on those verbs prove to be useful in the selection of specific information, identified on the basis of the concepts modelled in the *Privacy Policy Personal Data* CP. Those lexico-syntactic patterns are particularly effective in identifying the agents that are involved in a processing scenario, as it is shown in Table 5.6. Among the 29 template components that refer to an Agent and that were used to compute the accuracy in this step, 28 of them received a positive assessment from both experts. Moreover, considering individually the performance of the system in associating sentence excerpts to the *personal data* and the *purpose* components of the processing templates, the precision of the system is equal, in both cases, to 66%. This result suggests that, regardless the mappings drawn with specific concepts in the DPV modules, the combination of text chunks extracted by the vocabulary-driven approach and the information about optional constituents and co-occurrences of clauses in a sentence, contribute in the detection of sentence excerpts that are semantically related to the type of information represented by a template component.

Finally, the third evaluation objective **(o3)** concerns the assessment of the precision of the mappings between sentence excerpts and concepts in the DPV. The experts' evaluation referring to this objective was investigated by question Q3. By looking at the last four rows in Table 5.6, the recorded data show that, among the 131 template components that passed the second level of evaluation, 100 of them were positively assessed by both experts at the third level, resulting in a precision of 73%, that decreases to 52.1% when it is computed with respect to the 235 template components extracted by the system. Looking at the specific template components, it can be noticed that, also in this last step of the evaluation, the experts agreed in positively assessing the mappings between the verbs used to identify processing scenarios in sentences and the corresponding concepts in the DPV *Processing* module. This result suggests that the mappings that were first envisioned as lexical matches between verbs and class names could be turned into semantic ones.

When considering the mappings between text chunks and concepts in the DPV *Personal Data Category* module, the precision of the system is equal to 64%, measured on the *personal data* template components that passed the second level of evaluation. This percentage decreases to 42.1%, when considering the overall number of *personal data* template components under evaluation. Similarly, considering the mappings between text chunks and concepts in the *Purpose* module, the precision is equal to 43.8%, that decreases to 23.3% when considering the overall 60 mappings under evaluation. Those precision values are higher than those measured when the vocabulary-driven extraction of concepts was first tested as stand-alone approach applied to the OPP-115 corpus (see Section 4.2.4). Despite that evaluation was partial and based solely on some correspondences between the annotation in the OPP-115 and the concepts in the DPV, the increase of precision of the mappings for both *Personal-DataCategory* and *Purpose* modules, suggests that the incorporation of information about the sentences' clauses can improve the extraction results of the vocabulary-driven approach, filtering out noisy extractions.

5.3.4.1 Discussion of the Results

The results obtained from the experts' evaluation can be discussed in the light of the research question investigated in the thesis and the applicability of the approach with respect to the state of the art in the field of knowledge representation and NLP for the data protection domain.

The starting point for this thesis work was the definition of ODP, described as a "*modeling solution[s] to solve a recurrent ontology design problem*" [82]. Considering this definition, I assumed that if an ODP should represent a recurrent ontology design problem, then evidence of this recurrence should be retrieved in the texts belonging to the domain of interest modelled by the pattern. Having considered the data protection field as the domain of interest in this thesis, the research question derived from the formulated hypothesis and aimed to investigate the possibility to detect recurrent information scenarios for which a solution provided by an existing ODP already exists.

Based on this concept of *recurrence*, I primarily investigated the possibility of detecting the processing scenario of the *Privacy Policy Personal Data* pattern within the sentences of the privacy policies. The implementation of the system that tries to achieve this goal is based on the analysis of frequent lexico-syntactic patterns in the text of privacy policies, taking advantage of the information modelled in the DPV for the identification of those patterns. In computational linguistics, measures that rely on the calculation of term frequency for determining the relevance of terms have been traditionally criticised, leading to the introduction of more sophisticated variants of this measures (see Section 4.1.2). However, some distinctive features in the communicative style used in privacy policies, can help in identifying recurrent information scenarios within those documents. Moreover, another explanation that justifies how the reliance on the frequency of terms in privacy policies can be effective for the system concerns the specificity of the terms for which this frequency is calculated. In fact, the verbs that are modelled in the *Processing* module of the DPV directly recall the terminology used in Art. 4(2) of the GDPR, which defines the term *processing* through the enumeration of various processing activities. The fact that the system relies on those specific terms to select phrases describing processing scenarios and that the system performance achieved an 88% in accuracy shows that these terms are also frequently used within privacy policies. Consequently, the analysis of terms frequency was done in a "controlled" manner, relying on a set of domain-specific terminology, rather than an heterogeneous set of domain-independent terms.

Moreover, the regularities in the text that are captured by the implemented approach revealed the presence of another recurrent information in the text of privacy policies, i.e. the information about the roles played by different agents involved in a processing scenario. Indeed, driven by the information modelled in the *Privacy Policy Personal Data* pattern, the proposed approach is able to find in the text mentions to agents that participate in data collection and disclosure activities. Consequently, for those sentence excerpts, an extension of the method could try to map those mentions to agents on the corresponding roles that they play in the processing scenario. This information could be modelled through the *Agent Role* ODP (that was provided as an example in Section 2.2.3), relying on the concepts provided in the DPV to represent the agents' roles involved in a data processing activity (see Figure 4.3). Another proof of the recurrence of this informative scenario is implicitly provided by the fact that the SVO clauses, used to select the sentences in privacy policies, are those in which the subject corresponds to the pronoun *we*. Consequently, each occurrence of that pronoun could be mapped on the corresponding role of data controller, played by the company that discloses the privacy policy. Therefore, such a type of analysis of the privacy policies text with respect to existing ODPs could support the expression of the implicit information enclosed in the text, in order to support mechanism of automatic reasoning applied on privacy policies.

Another advantage of a system that is based on the information modelled in existing ODPs for extracting the information in privacy policies is the reusability of its outcome. Indeed, an information that refers to the concepts modelled on existing ODPs could represent a middle-layer of processing that fills the gap between an unstructured text and the formal representations of knowledge provided by different ontologies. Indeed, several ontologies that have been proposed for modelling the data protection field (see Section 2.3.1.4), could benefit from the information extracted by the implemented system. For instance, the concept that refers to the purpose of the processing is modelled in the ontology proposed by Bartolini [15], in the PrOnto ontology [150] and in the GDPRov ontology [154]. Similarly, the representation of the agent involved in a data processing activity is modelled by PrOnto, GDPRtEXT [153], and GDPRov. Consequently, despite the method was designed independently by the different knowledge models provided by the those ontologies, the information extracted could be reused for finding instances of specific concepts modelled in different ontologies. Moreover, several approaches have manifested their interest in the DPV and their intent to include it in their projects [108, 25, 169, 53]. However, no approaches for the automatic detection of concepts in the text of privacy policies have been proposed yet. Consequently, this system could fuel further investigation about the applicability and the customisation of the proposed approach to the requirements of specific research projects.

Among the weaknesses of the system, it could be pointed out its focus on single sentences for extracting the information. Indeed, the information that characterises a processing scenario could be spread over several sentences. In the following example, taken from the *mydramalist.com* privacy policy:

> *"We may determine the approximate location of your device from your IP address. We use this information to calculate how many people visit our Services from certain geographic regions."*

The system could detect a mention to a processing scenario in the second sentence, based on the presence of the verb *use* with the subject pronoun *we*. From this sentence the system could extract the purpose of the processing activity, but fails to associate the specific information about the personal data being used, mentioned in the first sentence, to the processing scenario detected from the second sentence. Consequently, the implementation of the system could be improved with a mechanism of co-reference resolution to enlarge the detection of the characteristics of a processing scenario from different sentences. Moreover, the verb *determine* used in the first sentence, highlights how data processing activities could be represented by verbs other than those represented in the DPV and used by the system to select the sentences. In the example, in fact, the verb *determine* represents a data collection activity, but would not be detected as such by the system. An evaluation of the information loss due to the constraints set by the implemented method to extract mentions to data processing activities has not been considered, because it would have required the experts' annotation of all the sentences in the corpus and the workload would have been excessive for them, considering their participation on a voluntary basis, without being remunerated for their work.

Another weakness that emerges from the expert evaluation concerns, more specifically, the vocabulary-driven approach. While some weaknesses were noticed in the development of the approach and were addressed in the implementation of the integrated approach for the detection of processing scenarios, some further improvement is still needed. The reliance on matches between terms in text chunks and the terminology in the DPV is in some cases misleading (see Section 4.2.4). For instance, a noun chunk *marketing communication*, that is used in privacy policies to refer to the purpose of the processing, is mapped by the system on the *Communication* class, which represents a type of personal data. This mapping is drawn because of the lexical match of the word *communication* with the homonym concept in the DPV *Personal Data Category* module . Moreover, because the text chunk appeared in the sentence with other two text chunks mapped on concepts of the *Personal Data Category*, according to the implementation of the system, the text chunk was erroneously considered as a correct mention to a personal data. To improve the performance of the system in such situation, the similarity values that are computed by the vocabulary-driven approach to associate a text chunk to a concept could be further exploited to filter out other noisy extractions.

5.4 Summary

This chapter presented the evaluation of the integrated approach for the extraction of processing scenarios from the text of privacy policies. The system was tested on a GDPR-oriented corpus of privacy policies, selected from the Princeton-Leuven Longitudinal corpus. The evaluation relied on the assessment of the results by two legal experts. The provided annotation proved to be reliable considering the annotator-agreement computed with the Cohen's κ coefficient. The experts' evaluation showed that the system can detect sentences that describe a processing scenario with a precision equal to 88%. The integration of information about the clauses of the sentences and the mappings with some DPV concepts has proved to be useful in the characterisation of the processing scenarios with respect to the concepts modelled in the *Privacy Policy Personal Data* pattern.

6 Related work

This chapter provides an analysis of some works that are related to the topics covered by this thesis. The first part presents the approaches to the use of ODPs in literature and the automated approaches to the processing of legal texts in legal fields other than the data protection one.

The second part of the chapter focuses on related works in the data protection field, trying to highlight similarities and differences of the existing approaches and the work presented in this thesis.

6.1 Ontology Design Patterns in Literature

The most common application of ODPs in literature has concerned their adoption in the design of modular ontologies, as envisioned when they were first proposed. The ontologies whose design has been based on the reuse of existing ODPs are mostly domain ontologies, scoping a variety of fields.

When describing the feature-based analysis of existing legal ontologies in Chapter 2, I already mentioned the PrOnto ontology, that has been proposed to model the data protection field in order to support automated approaches to compliance checking with the GDPR. The modular structure of the ontology relies on the *Time-interval* pattern to represent intervals of time and the *Time-indexed Value in Context* [1] to represent the scope and the temporal interval in which an entity assumes a specific value.

Another ontology which is based on the reuse of existing ODPs has been proposed by Elhassouni et al. [62], for modelling credit risk scorecard and supporting decision-making processes in the financial field. The ontology reuses and specialises the *Event* pattern to model the credit risk scorecard and the *Agent Role* CP for representing credit risk scorecard players. Given the representation of the event risk scorecard as an event, its variables (e.g. age of the individual applying for credit, account balance, guarantor) are represented as

[1] `https://sparontologies.github.io/tvc/current/tvc.html`

participants in the event. Finally, the *Classification* CP is used to model different categories of variables and the credit risk scores (i.e. low, medium, high).

A relevant effort in reusing existing ODPs in ontology engineering is made by the ArCo ontology [31], that addresses the modelling of the Cultural Heritage Domain. The ArCo project started with the aim of converting the information stored in the relational database of the General Catalogue of Italian Cultural Heritage in a Semantic Web oriented knowledge graph. The ArCo ontology reuses existing ODPs for modelling several scenarios, such as tracking changes made to a catalogue record that describes a cultural property. In this case, the *Information Relization* and the *Time Interval* CPs are jointly used to model a catalogue record and the interval of time of its validity. Then, groups of various versions of a catalogue record are represented using the modelling solution provided by the *Sequence* CP. ArCo also models the different places where a cultural property could be located over its life and the situations in which it could be involved (e.g. commission, trade, obtainment). The representation of those scenarios rely on the specialisation the *Time Indexed Situation* and the *Situation* CPs. In the development of the ontology, an new CP, named *Recurrent Event Series*, has also been proposed and submitted to the ODP portal[2]. The pattern models recurrent events, i.e. those happening at regular time intervals, by modelling each occurrence of an event through the *Situation* CP and modelling different occurrences though the *Sequence* CP.

Another application of ODPs in the ontology design process was investigated by Aguado de Cea et al. [3] for automatising the choice of suitable ODPs for fulfilling specific modelling requirements. Specifically, the authors envisioned the development of the S.O. S. (System for Ontology design pattern Support) tool able to semi-automatically propose a set of ODPs starting from the formulation in natural language of a modelling requirement. The core of the system lies on some lexico-syntactic patterns composed by *subject-verb-object* triples able to unravel semantic relations expressed in the formulation of the modelling requirement and corresponding to the relationships modelled by different ODPs, e.g. *subclass-of*, *equivalence*, *part-whole* or *participation* relations. The authors also considered situations of non-unique correspondence between a lexico-syntactic pattern found in the natural language formulation of the modelling requirement and a relationship in the ODPs. On the one hand, this situation could be due to the intrinsic polysemy of verbs used in the lexico-syntactic patterns, that could express different types of relations (e.g. the verb *include* could be equally used to express a *subclass-of* and a *part-whole* relation). On the other hand, the non-unique correspondence between a lexico-

[2] http://ontologydesignpatterns.org/wiki/Submissions:RecurrentEventSeries

syntactic pattern and ODPs relations could be rise from the need of using more than one ODP to fully realise the pattern. The S.O.S. system has some similarities with the work that I presented in this thesis. First, it is comparable, in its objective, to the InvetigatiOnt tool that I implemented and described at the end of Chapter 2. Indeed, both systems aim to encourage and ease the reuse of existing ontological models. However, while the S.O.S. system focuses on existing general-purpose ODPs and the relations that they model, the InvestigatiOnt tool focuses on a more conceptual level, aimed at discovering and understanding the ontological commitment of different legal ontologies. Second, the S.O.S system is comparable to the work described in this thesis in its reliance on lexico-syntactic *subject-verb-object* patterns. However, while the patterns used in the S.O.S. system are crafted to discover more generic domain-independent relations, the lexico-syntactic patterns that were used by the system described in Chapter 4 were tailored on a specific CP addressing the data protection domain.

6.2 Approaches to Automated Processing of Legal Texts

The application of OIE techniques on legal documents has been investigated by Siragusa et al. [180] who proposed the LegOIE system. The implementation relies on the Inter-Active Terminology for Europe (IATE), i.e. the EU's terminology database. In a preliminary step, the terminology in IATE is filtered out of noisy terms and multi-word-expressions that are not associated with a domain label or that have a scarce correlation with it. To find the terminology that has a low correlation with its domain label, the authors trained a word-embedding model on a corpus of European Directives and Statutory Instruments documents. The word embedding were used to filtered out the *(terminology, domain)* pairs with a cosine similarity value lower than a given threshold. Based on this preliminary step, LegOIE processes sentences from legal documents for extracting relational phrases that link couples of terms or multi-word-expressions that appear in the filtered version of IATE. To find those relational phrases, the system first generates the dependency graph of the sentence, merging the words in the graph that form a single IATE concept. Afterwards, considering the undirected version of the graph, it extracts the shortest path that connects two IATE concepts in the sentence, considering as valid relational phrases only those paths that contain a verb. Each extracted triple, having the form *(iate_concept_1, relational phrase, iate_concept_2)*, is associated with a confidence score that is proportional to the frequency of the triple in a corpus of documents multiplied by the inverse of the relational phase's length.

The scores are used to order the triples by decreasing confidence value. The system is tested on a corpus of 4310 European Directives, from which LegOIE extracted 2267 triples. The performance of the system is compared with the performance of ReVerb, OllIE and ClausIE. Based on the manual evaluation of a sample of 100 triples extracted by each system, the LegOIE reached the best accuracy value, being equal to 0.32. No mention to the precision of the results is provided.

The IATE database was also used in the concept recognition system implemented by Nanda at al. [136]. In their approach, the authors automatically annotated a corpus of European directives and national law from United Kingdom, by looking in the corpus for those terms that matched a IATE entry and eventually associating them with the corresponding domain label. Moreover, an existing Named Entity Recognition system was used to label concepts in the corpus representing time, date and monetary units. The 80% of the annotated corpus was used to train the concept recognition system, based on conditional random fields utilizing word suffix, word identity (i.e. whether a word represents a subject domain/named-entity or not), word shape (captialized, lowercase or numeric) and part-of-speech tags as features. When performed on the test set, the system achieved a precision equal to 0.76 and a recall equal to 0.68, resulting in a F1 score equal to 0.71.

The two approaches described above, although not explicitly stated by the authors, could be potentially used to learn concepts and relations in the ontology learning task. By contrast, an approach that was specifically envisioned for learning legal ontology components is that proposed by Lame [110], that specifically focused on concepts and relations learning from a corpus of 57 Codes in French Law. He first populated a list of candidate concepts, identified by nouns and noun phrases extracted utilising an existing syntactic analyser. To discriminated between legal and non-legal concepts in the list, the author investigated different classical statistical methods for weighing index terms (i.e. frequency, TF-IDF and entropy). However, he concluded that none of them was reliable for the identification of legal concepts, but, instead they could have been used for cleaning the list of candidate terms from "empty terms", such as *article* or *chapter*. Thus, he relied on frequency values to perform this filtering phase. From the list of remaining candidate concepts, he identified a list of fundamental legal terms by using discourse structures, that analyse specific parts of the text, such as titles and summaries, for extracting core concepts in a domain of interest. For identifying couples of related legal concepts, he applied a statistical method which analyses the context words surrounding the identified legal concepts, based on the assumption that similar concepts share similar semantic contexts. Each legal concept was associated to a vector of context words, each of them weighted with a mutual information score that quantifies the dependency in the corpus between the context word and the legal concept. The mutual information score was computed taking into account both the joint frequency of the words and their individual frequencies. Given such a vector representation, the cosine similarity is used to find the most related pairs of concepts. Finally, the specific relations holding among those pairs of related concepts was inferred manually from the analysis of the concepts.

A different approach to information extraction for populating legal ontologies is described by Humphreys et. al. [102], who combined a rule-based approach and the Mate tool for SRL (see Section 4.1.1) to extract definitions, norms and their elements from legislative texts. First, based on some lexico-syntactic patterns, the system distinguish between sentences that represent definitions and sentences that represent norms. Second, different roles are looked at in the sentences, according to the detected distinction between definitional and normative sentences. For definitions, the semantic roles identified by the system are: Definiendum, Definiens, Includes and Excludes. Those roles are identified by manually establishing some mappings between the semantic roles identified by the Mate tool and the legal roles expected in definitional sentences. By contrast, the roles extracted by the system for a normative sentence depend on the detected type of the norm, that is identified by analysing the head verb of the sentence. The types of norms that the system recognises are: Definition, Obligation, Permission, Power, Scope, Right, Hierarchy, Exception and Legal Effect. For instance, when the detected norm type is Obligation, Permission, Power or Right, the system looks for the following semantic roles: Action, Active Role, Passive Role, Condition, Timeframe, Exception and Reason. Similarly to the approach followed in the detection of roles in definitional sentences, the detection of roles in the different type of norms relies on hand-crafted rules that map the general roles extracted by Mate with the domain-specific legal roles. The performance of the system was tested on a EU directive. Precision, Recall and F-measure of the system are computed in two evaluation settings: strict evaluation takes partially correct results as wrong, whereas lenient evaluation consider them as being correct. The system achieves good F-measure value in detecting the norm type, both in the strict and the lenient setting (81.6%). A similar result (around 80% in both evaluation settings) is achieved when extracting the Active Role. By contrast the performance of the system on detecting other roles is variable and asks for future improvements, as claimed by the authors.

The methods described so far applied different techniques for extracting concepts and relations that could be exploited to partly automatise the process of ontology building for the legal domain. The approach proposed by Buey [28], instead, investigated the opposite approach, i.e. the reliance on an existing ontology to boost automated information extraction from legal texts. The proposed approach is designed to deal with four types of legal documents, i.e. notarial acts, judicial acts, registry documents, and private documents, which vary in their level of structuring and the type of information that they convey. The implemented information extraction process is based on the information modelled in an ontology, which stores information about the text structuring and the types of entities that are expected to be mentioned in the different document types. A preliminary pre-processing step corrects misspelled words and removes noisy terms, such as signatures and stamps, that could hinder the performance of the system. Knowing the type of the document, its cleaned text is processed to identify the different sections it is made of, based on the information stored in the ontology, that also provides the system with information about the technique to be invoked to detect those sections. Having identified the text sections with the corresponding text paragraphs, the last step of the method extracts the entities mentioned in them, togeter with their properties. To perform this step, the system consults, again, the ontology that provides it with information about the types of entities to be extracted and the method to be applied for executing the extraction. Thus, the proposed approach is conceived as a framework that integrates in a single solutions, the inevitably heterogeneous methods that must be applied to handle the text processing and the entities extraction. The system has been tested on a corpus of 144 Spanish notary acts. The ontology used to guide the extraction contained information about two approaches for segmenting the texts in sections and 17 rule-based approaches for extracting entities. The overall performance of the system achieves a F-measure equal to 80%.

6.3 Automatic approaches to GDPR compliance checking

Many European projects have addressed the challenge of implementing automated approaches to develop services of automatic compliance checking with the Regulation. A brief overview of those projects could help to highlight how the work presented in this thesis could potentially be adopted to support wider projects (see Chapter 7 for further discussion).

Claudette[3] [42] exploits machine learning and grammar based approaches to provide consumers with a tool able to automatically detect potentially unfair

[3] http://claudette.eui.eu/index.html

clauses in online terms of service. After the good results reached in the consumer law domain, the focus of the project has now shifted on the analysis and evaluation of privacy policies along three dimensions: comprehensiveness of information, substantive compliance and clarity of expression.

The SPECIAL[4] project [25] focused on the development of machine-readable policy languages for expressing consent, business policies and regulatory obligations. Based on these languages, it implemented reasoning algorithm to automatically check if a business process complies with the consent given by the data subject and the obligations set by the GDPR. The DPV has been released in the context of this project.

The MIREL[5] project focused on the representation of legal norms and the implementation reasoning mechanism based on ontological representations of legal concepts to support compliance checking in the data protection domain [149]. The aforementioned PrOnto ontology has been released in the context of this project. The PrOnto ontology has also been used in the DAPRECO project [167] to released the DAPRECO knowledge base, that provides a machine-readable representation of the norms in the GDPR. The norms are represented in reified Input/Output logic and encoded in LegalRuleML.

The Lynx[6] project aimed to create a legal knowledge graph to manage compliance with the law. One of the pilots designed to test the project approach concerned the data protection field and its goal is to create a knowledge graph for the data protection field interlinking domain-related legal texts and providing algorithms able to automatically enlarge the knowledge base when new relevant documents are issued [132].

The SMOOTH project[7] aims to assist micro enterprises to become compliant with the GDPR by designing and implementing tools for the validation of compliance according to the existing legislation. The SMOOTH platform will be built upon several existing techniques, combining advanced technologies in the area of machine learning, text mining and data mining.

6.4 Approaches involving privacy policies

This section presents the NLP approaches that has been implemented to process the text of the privacy policies, shaping similarities and differences with the work presented in this thesis.

[4] https://www.specialprivacy.eu/
[5] https://mirelproject.eu/index.html
[6] http://www.lynx-project.eu/
[7] https://smoothplatform.eu/

6.4.1 Classification of privacy policies' paragraphs (with supervised models)

6.4.1.1 Polisis: Automated Analysis and Presentation of Privacy Policies Using Deep Learning

Polisis is a framework used for implementing a question-answering system that selects the most suitable paragraphs of a privacy policy for answering the queries formulated by users about the processing of personal data performed by a company. The core of Polisis is a machine learning module made of convolutional neural networks classifiers which leverage privacy-specific word embeddings, built from a corpus of mobile apps' privacy policies. The training of the classifiers relies on the OPP-115 corpus to predict the labels at paragraph and text span level, both in the privacy policies and the queries formulated by the users. The labels are used to compute a proximity score between a query and each paragraph of a privacy policy to select those that can answer the query. The performance of the question answering-system is evaluated building a dataset of questions that Twitter users addressed to companies. The dataset is used, then, to compute the accuracy in the answers provided by the system with those manually selected by two annotators.

Similarities: The use of the OPP-115 corpus.
Differences: The OPP-115 corpus is used to perform a classification task that supports a question answering system.

6.4.1.2 Establishing a strong baseline for privacy policy classification

In their work, Nejad et al. [141] pinpoint a lack of information in the description of the Polisis framework (see Section 6.4.1.1), that prevents the reproducibility of the machine learning module implemented to train the convolutional neural networks classifiers. Consequently, the authors address the task of providing a baseline for the automatic classification of privacy policies using the OPP-115 corpus as a dataset and stressing on the reproducibility of their experiments. Their classification task is implemented comparing three models: a convolutional neural network that relies on privacy-specific word embeddings, a pretrained version of the BERT framework and a fine-tuned version of BERT that relies on a large corpus of mobile apps privacy policies. The evaluation of those models is made considering two gold standards derived from the OPP-115 corpus, considering the majority vote and the union of experts annotations respectively. The classifiers performances are provided measuring both micro-average and macro-average F1, demonstrating the dominance of both the BERT settings compared to the results obtained by the convolutional neural network.

Similarities: The use of the OPP-115 corpus. **Differences:** The OPP-115 corpus is used to perform a classification task.

6.4.1.3 Towards Measuring Risk Factors in Privacy Policies

The framework described in [141] constitutes the base for a theoretical pipeline described in Nejad et al. [140]. In their proposal, they envision an architecture for the detection of risk factors in privacy policies. The first part of the pipeline is constituted by the classification framework described in [141] that assigns high level labels taken form the OPP-115 corpus to the privacy policies paragraphs. In the following step of the pipeline, the high-level classification of the paragraphs is refined by a rule-based approach that extracts the values associated to the attributes define in the OPP-115 corpus. The approach should rely on some lexico-syntactic patterns manually defined by experts, but no further details are provide about the implementation of this step. Based on the attribute-value pairs detected with the rule-based approach, the last module of the pipeline assigns a risk level to each implemented data practice detected in a privacy policy. Because the proposed pipeline is described at a theoretical level (except for its first step) no evaluation is provided to assess its quality.

Similarities: The use of the OPP-115 corpus and the reliance of lexico-syntactic patterns for the extraction of fine-grained attributes from the privacy policies. **Differences:** The lexico-syntactic patterns are manually defined, while the approach proposed in this thesis applies OIE to discover those patterns automatically.

6.4.2 Topic-modelling (unsupervised models)

6.4.2.1 Unsupervised topic extraction from privacy policies

This work [171] proposes a semi-automated framework for the detection of relevant topics in the privacy policies paragraphs. The proposed approach applies the Latent Dirichlet Allocation (LDA) model [22] to extract 100 topics from a corpus of 4982 privacy policies crawled from the Web. A further phase of expert manual analysis of the extracted topics is performed to discard non-cohesive topics and to merge similar topics. The result of this manual processing is a set of 36 topics to describe the content of the privacy policies paragraphs.

The validation of the proposed framework relies on the OPP-115 corpus and show, through a manual mapping, how the final set of topics overlaps the labels provided by the corpus at paragraph and fragment level.

Similarities: The use of the OPP-115 Corpus and the validation based on a manual study of the overlap between the corpus labels and the topics.
Differences: The topics are not organised in a taxonomy or in any other specified semantic structure.

6.4.3 Question-answering over privacy policies

6.4.3.1 RECIPE: Applying Open Domain Question Answering to Privacy Policies

The RECIPE [176] methodology grounds on the theory of contextual integrity. Based on this theory, the description of the flow of personal information should specify five parameters that include the type of information and its subject, the sender and the recipient of the information, and the conditions under which the data flow occurs [144]. The RECIPE methodology builds five questions on such parameters and combines two approaches to answer those questions with respect to the information disclosed in privacy policies. First, a pretrained model for open domain question answering is applied to extract the parameters at paragraph level. Second, the parameters of an information flow are detected by applying a dependency parser at sentence level and establishing some mappings between the dependency types and the flow parameters expressed by such dependency types. The outputs of the two approaches are manually checked and merged to answer the five questions referring to the parameters that characterise the description of a personal information flow.

The RECIPE methodology is evaluated by computing the F1 score with respect to six manually annotated privacy policies from the OPP-115 corpus. The results show that the F1 score achieved by combining the open domain question answering model and the dependency parsing approach outperform the F1 scores achieved by the individual approaches. The authors do not mention the values of precision and recall that were used to compute the F1 scores.

Similarities: The information described by the contextual parameters is similar to some classes in the *Privacy Policy Personal Data* ODP. The OPP-115 corpus is used by limiting the experiments to the paragraphs having the labels: *First Party Collection/Use, Third Party Sharing/Collection, Data Retention*.
Differences: the contextual parameters are nor organised in some semantic superstructure.

6.4.4 Mapping between privacy policies and laws

6.4.4.1 'KnIGHT: Mapping Privacy Policies to GDPR"

The KnIGHT tool [142] uses semantic text matching for mapping the relevant sentences in the privacy policies' text to the most related article and paragraph of the GDPR.

The software architecture of KnIGHT relies on two processing steps. The first step extracts salient terms both from a corpus of twenty privacy policies of EU-based companies and from the set of ninety-nine articles of the GDPR. The second processing step implements a semantic text matching algorithm to compute the similarity between the set of salient terms in a privacy policy's sentence and the set of salient terms for each of the ninety-nine articles of the GDPR. The two sets of relevant terms are represented as two word embedding vectors and their similarity is computed using the cosine similarity measure. The GDPR article that produces the greater similarity score is considered as the most relevant with respect to the privacy policy sentence. Following a similar approach, the most relevant paragraph of the selected article is identified encoding both the privacy policy's sentence and the paragraph's article as the average of their word vectors. Then the cosine similarity measure is applied to select the paragraph with the highest similarity score.

The evaluation of the tool is provided as a posteriori assessment where four legal experts evaluated the output of KnIGHT applied on a set of four privacy policies. According to the proposed evaluation a variable rating spanning from 70% to 90% of the mappings found by KNIGHT are at least partially correct.

Similarities: The use of a text similarity measure.
Differences: The objective of the experiment: KnIGHT maps the privacy policies sentences with the text of the Regulation, whereas this thesis uses the similarity measures to perform mappings between phrases in privacy policies and concepts in the DPV vocabulary. Moreover, the relevant terminology is identified in Knight on a statistical analysis of the corpus, while the thesis project adopts a joint approach drive both by statistical analysis and by the DPV.

6.4.5 Summarisation of privacy policies

6.4.5.1 PrivacyCheck: Automatic Summarization of Privacy Policies Using Data Mining

PrivacyCheck [199] is a browser extension that automatically processes the text of privacy policies to summarise them through graphical icons that highlight the risk factors for users' privacy. PrivacyCheck identifies ten risk factors and applies supervised machine learning models to detect those factors in privacy policies text. Specifically, a model is trained to recognise whether the Web page specified by an URL is a privacy policy and other ten different models are trained to recognise each of the risk factors. No further details are provided by the authors about the adopted machine learning models whose implementation is delegated to the Google Prediction service. The training set includes 400 privacy policies of companies selected from the lists provided by three American stock markets. The documents were manually annotated to identify the risk factors. The performance of the classification models is assessed computing the F1 score on a test dataset of 50 privacy policies manually annotated following the same criteria used to label the training dataset. The authors do not mention the precision and recall values that were used to compute the F1 value.

Similarities: -
Differences: PrivacyCheck is different from the research proposed in this thesis both in its scope and implementation.

6.4.5.2 PrivacyGuide: Towards an Implementation of the EU GDPR on Internet Privacy Policy Evaluation

PrivacyGuide [187] provides a visual summarisation of the privacy policies content, highlighting the risks associated to some privacy-related aspects described in those documents. The system identifies eleven privacy aspects that result from the interpretation of the GDPR made by some legal experts. Each privacy aspect can be associated to a risk level in a three scale, depending on the information provided by a privacy policy sentence. To recognise the privacy aspects referred to in the privacy policies sentences, eleven classifiers are trained on a corpus of 45 manually labelled privacy policies using TF-IDF vectors to represent the input documents. The results of four classifiers (Naive Bayes, SVM, Decision Tree and Random Forest) are compared to measure precision, recall and F1. Considering those metrics, the Naive Bayes model is selected to build the final model at the core of PrivacyGuide. The output probabilities of the model are, then, used to associate to each privacy aspect the corresponding risk level. The average precision reached by the system is equal to 68.4%.

Similarities: Experiments are based on a GDPR-compliant corpus, collected from Alexa rankings.
Differences: The goal of the experiment is the summarisation of privacy policies with privacy risk assessment.

6.4.5.3 Toward Domain-Guided Controllable Summarization of Privacy Policies

In their work, Keymanesh et al. [107] implement an extractive approach to generate a summary of privacy policies by selecting those sentences which could reveal a high risk for the privacy of users. The proposed method relies on pre-trained ELMo word embeddings [159] and a convolutional neural network classifier that predicts the risk probability associated to each sentence. The summary of a privacy policy is, then, obtained with two alternative approaches. A risk-focused approach selects the sentences associated with the highest probabilities scores. By contrast, a coverage-focused approach identifies clusters of sentences that refer to the same risk factor, selecting the sentence with the highest risk probability in each cluster. The experiments are implemented using a corpus of 151 manually labelled privacy policies from the TOS;DR[8] corpus. The performance of the classifier is assessed computing the F1 metric with respect to the corpus' manual annotations. Moreover, the quality of the generated summaries is compared to that of different domain-independent baseline models computing standard metrics for text summarisation tasks (i.e. ROUGE and METEOR).

Similarities: -
Differences: The approach is different from the research proposed in this thesis both in its scope and implementation.

6.4.6 Open Information Extraction for Ontology Refinement

6.4.6.1 Hybrid Refining Approach of PrOnto Ontology

In this work [147], two legal experts annotated a corpus of 10 privacy policies with labels that represent different concepts relevant for the GDPR. Moreover, the concepts modelled in PrOnto (the ontology was described in Section 2.3.1.4 of Chapter 2) were manually associated to different lexical variants found in privacy policies. The annotated privacy policies, together with the text of the GDPR were used as an input for an OIE system that maps sentence excerpts of privacy policies on the concepts of the PrOnto ontology. The

[8] https://tosdr.org/

extracted mappings were analysed by the legal experts to select the correct and the incorrect ones. Then, the assessment provided by the experts was provided, again, in input to the OIE system for refining its extraction. This iterative process was performed three times and the final output of the system was used to refine the PrOnto ontology. The recall of the implemented system is reported to be equal to 75% in detecting excerpts from privacy policies that could be mapped on the classes of the PrOnto ontology. The authors do not mention the precision of the implemented system.

Similarities: The reliance on a technique and the concepts an ontology to drive the extraction of information from privacy policies
Differences: The approach relies on the manual annotation of lexical variants of the concepts modelled in the ontology. By contrast, the approach proposed in this thesis makes an effort to detect those lexical variants automatically, relying on manual annotation only for the final assessment of the system.

6.5 Summary

This chapter presented the state-of-the-art work related to this thesis. The first part of the chapter presented some works that used ODPs mainly for modelling new ontologies, demonstrating that ODPs are actively used in the research community that works on knowledge representation topics. Moreover, several approaches for the processing of legal texts in domains other than the data protection field were presented.

The second part of the chapter presented the related works in the data protection field, highlighting similarities and differences with the work proposed in this thesis. Despite some commonalities in the resources adopted for performing the experiments, the approach proves to be different from many existing ones.

7 | Conclusion and Future Work

This chapter ends the presentation of the thesis, with a summary of the methodology implemented to investigate the research question, shaping the direction for future work.

7.1 Summary of the Research

This thesis started from the definition of Ontology Design Pattern provided by Gangemi and Presutti [82], who described it as a "modelling solution to solve a recurrent ontology design problem". This definition was put into perspective in the data protection field, where the efforts in formalising legal knowledge aim to support the implementation of services for automatic compliance checking and a transparent processing of personal data. In order for those systems to work on real use cases, ontologies could benefit from the information extracted from privacy policies by means of NLP techniques. However, at present, most existing approaches to the automated processing of those documents have not considered this opportunity.

The intuition that led to this work was the possibility of relying on ODPs as a means to reconcile the multitude of ontologies that have been released to model the data protection field and the disparate approaches that have been proposed to automatically process the text of privacy policies. Based on this intuition, the thesis investigated the possibility of extracting from the text of privacy policies the recurrent informative scenarios for which a modelling solution already exists. To find those scenarios, the proposed solution combined NLP techniques and the information provided by the DPV. The results showed that at least three ODPs (i.e. the *Privacy Policy Personal Data*, *Agent Role* and *Collection* patterns) could be exploited to model the information in privacy policies.

7.1.1 Adopted Methodology

The research grounded in a wide comparative study of existing ontologies that have been proposed in the last decade to model different legal fields. This analysis was presented in Chapter 2 and highlighted the variety of possible representations of legal knowledge. The first outcome of this analysis was the development of a Web application that helps users interested in the reuse of existing ontologies to explore the variety of those knowledge models. The second outcome of the analysis was the acknowledgement of the increasing interest in modelling formal representations of knowledge in the data protection field, that was the domain in which most ontologies were found. Many of them considered the legal framework set by the GDPR, that was presented in Chapter 3, together with a discussion of the critical aspects of privacy policies which undermine a transparent communication of the information to be provided to individuals when their personal data are processed (Art. 13 and Art. 14).

Privacy policies were the focus of Chapter 4, that explained the steps undertaken for automatically process their text in order to detect the information that can be organised semantically through existing ODPs. The implemented processing pipeline consisted of: *(i)* a preliminary manual analysis of existing ODPs for finding those of potential interest in the data protection domain, *(ii)* an OIE task for identifying recurrent lexico-syntactic patterns that unravel the presence of recurrent information in the text, *(iii)* an approach for extracting mentions to personal data types and purposes of data processing exploiting the concepts in the DPV and proposing the first solution based on the *Collection* ODP for mapping text chunks on concepts in the vocabulary, *(iv)* an approach that integrates the outcomes of the two previous steps to detect the information that could be modelled by the *Privacy Policy Personal Data* pattern. The performance of the system was evaluated by two legal experts and the results were described in Chapter 5. Based on the experts' assessment, the system has a precision of 88% in detecting the sentences that express an informative scenario corresponding to that modelled in the *Privacy Policy Personal Data* pattern. Moreover, the system detects specific sentence excerpts that correspond to some concept in the pattern with a precision of 71%, and to some concept in the DPV with a precision of 52.1%. The analysis of the results also highlighted the possibility to use another ODP, i.e. the *Agent Role* pattern, to model different stakeholders of a data processing activity. The implemented system is different in many ways from other existing approaches to automatic processing of the privacy policies' text. Those differences were discussed in Chapter 6 together with other related works.

7.1.2 Main Findings

Several considerations have emerged from this research. The legal domain is complex and can be formalised through ontologies that embrace different ontological commitments. ODPs offer various possibilities, both domain-specific and domain-independent, to provide a lightweight semantic structure to the unstructured text of privacy policies. Those patterns can be used as an interface between the text of privacy policies and the different ontologies that have been proposed in the data protection field.

However, the extraction of information from those documents is difficult and different types of NLP approaches may be needed for extracting different information. For instance, the implemented system revealed that not all types of information can be detected through the use of a single technique based on the detection of recurrent information. Moreover, the representation of the informative scenarios offered by the ODPs could be complex and the information in the text that refers to it could be scattered on different sentences and parts of the text, as revealed by the necessity of splitting the overall scenario modelled by the pattern in smaller parts.

Despite acknowledging that approaches of NLP on legal documents can not fully address some crucial aspects of the legal domain, such as legal interpretation, and that the formalisation of legal knowledge in highly-structured ontologies necessarily asks for human intervention, the proposed approach could assist the activity of legal and technological experts working in the knowledge engineering field. For instance, it may help to reduce the human effort in annotating documents with semantic metadata. Moreover, a system similar to the Web service that was presented in Chapter 2, could be extended providing a functionality for analysing the text of privacy policies, showing the available possibilities of modelling the information through existing ODPs.

The proposed approach could also be integrated in systems for monitoring compliance with the GDPR. In this respect, several works have expressed their interest in the use of the DPV [108, 25, 169, 53]. The approach presented in this thesis, in its effort to reuse as much as possible existing knowledge sources, represents the first NLP approach that relies on the information provided by this vocabulary for extracting information from privacy policies. Consequently, the mappings from the text to the concepts of the vocabulary could be integrated in those systems.

7.1.3 Future Work

The future work that follows from the research proposed in this thesis scopes different directions.

First of all, a future work may concern the enlargement of the mappings between text spans in the privacy policies and concepts in the DPV's modules that model the legal grounds and the agents involved in the data processing activities (see Section 4.2.1.3 in Chapter 4). Those mappings could be, thus, exploited both for refining the instantiation of the *Privacy Policy Personal Data* pattern (which models the *LawfulBasisForProcessing* concept) and populating another ODP that could be of interest in this domain, i.e. the *Agent Role* ODP, as previously discussed in Section 5.3.4.1 of Chapter 5.

Another future work may concern the application of co-reference resolution techniques. Indeed, the information related to a processing scenario could be scattered over multiple sentences. The approach presented in this thesis did not addressed this situation, however, co-reference resolution techniques could help to detect processing scenarios that are not fully described by a single sentence.

Moreover, while in this work I relied on word embeddings pre-trained on general-purpose corpora (see Section 4.2.4.4), existing words embeddings models could be retrained for learning domain-specific vector representations of words, taking advantage of the large availability of textual documents included in the Princeton-Leuven Longitudinal corpus. Among the existing approaches, the use of the BERT [57] language model is now the state-of-the-art approach in several NLP tasks [84, 134, 164]. Consequently, BERT could be used to learn a language model specific for the privacy domain. Alternatively, the adoption of existing BERT models pre-trained on corpora of legal documents could be investigated. This is the case, for instance, of the family of models provided by Legal BERT [35], which includes a model trained on the EU legislation.

For discovering textual evidence of other existing ODPs within the text of the privacy policies, a technique that should be prioritised is the named entity recognition. This technique has been applied in the legal field to the text of legislative documents, like in the approach proposed by Nanda et al. [136] (see Section 6.2). The identification of named entities, such as persons, organisations, locations or dates, could prove useful for discovering several ODPs, among those resulting from the analysis of the Ontology Design Patterns portal (see Section 4.2.2) and visually represented in Figure 4.5. Many of those ODPs model scenarios that involve the representation of temporal entities and agents, as it is the case for patterns like *TimeIndexedPaticipation*, *Participant-*

Role and *Action*. Consequently, the identification of the named entities within the text of the privacy policies, like the name of the data controller or the time limit for the retention of personal data, could help in populating other ODPs.

Another technique that should be investigated for discovering additional ODPs would take advantage of the recent progresses of neural networks models to perform classification tasks in a supervised setting. Those models have also been tested on the text of privacy policies, as in the case of some works presented in Chapter 6 (see, for instance, the implementation of Polisis, described in Section 6.4.1.1). Indeed, with a preliminary step of text classification of the privacy policies' paragraphs, the extraction of the information scenarios of interest could be targeted to the paragraphs in which, according to the predicted label,they are more likely to appear.

Concerning the evaluation of the method, a crowdsourcing approach hosted on dedicated platforms like Amazon Mechanical Turk[1] could be evaluated to extend the assessment of the results to a larger set of sentences. Moreover, the evaluation of the proposed approach could be adjusted to take into account the requirement of transparent communication set by the GDPR and the common hurdles that notoriously affect the transparency of information (both discussed in Chapter 3). Precision and, eventually, recall of the method could be interpreted taking into account these aspects. Thus, the possibility for the implemented method to exactly match the information described in one or more sentences with all the classes of an ODP could be a clue of a complete and well specified information. By contrast, the impossibility to identify the information scenario modelled by a patter could be an evidence of an information that is missing or vaguely expressed, hindering the possibility of an automatic approach to find it in the text. In the latter case, thus, the evaluation of the system should not be penalised.

[1] https://www.mturk.com/

Appendices

A Selection of Ontology Design Patterns from the portal

This Appendix presents the details of the iterative process aimed at selecting candidate content patterns of interest to the data protection field, as described in Section 4.2.2.

The following table lists, in the first column, the content patterns of the Ontology Design Patterns portal, in alphabetic order. For each pattern, the second column lists the domain labels associated to it, if any. The third column indicates the iteration that eventually filtered out a pattern from the final list of patterns of interest, according to the exclusion criterion formulated for that iteration (see Table 4.2). The forth column provides an explanation that justifies the discarding of a pattern made at the third and forth iteration of the process. In order to limit the number of rows of the Table, the patterns that were filtered out in the fist iteration of the process (i.e. those lacking of the competencies questions or the OWL building block) are not present in the table.

The rows highlighted in green indicates the patterns that have passed the four-stage elimination process and that were included in the final list of candidate patterns of interest to the data protection field.

CP Name	Domain	It.	Explanation
Acting For			
Action	Product development Businnes General		
Activity Specification	Event Processing General		
Actuation-Actuator-Effect	Internet of Things	2nd	
Affordance		4th	more related to physical objects
Agent Role	Management Organization Scheduling		

(Continue on the next page)

CP Name	Domain	It.	Explanation
Acquatic Resource Observation	Fishery	2nd	
Acquatic Resources	Fishery	2nd	
Bag	General Parts and Collections		
Born Digital Archives	Archives	3rd	too specific for being applicable in the domain of interest
Catch Record	Fishery	2nd	
Chess Game	Game	2nd	
City Resident Pattern	Smart City	3rd	too specific for being applicable in the domain of interest
Classification	General		
Climatic Zone	Fishery	2nd	
Co-participation	General		
Collection	General		
Collection Entity	Parts and Collections		
Componency	Parts and Collections	3rd	refers to physical objects
Computer System	General	3rd	refers to the software engineering domain
Constituency	Parts and Collections		
Course	Academy University	2nd	
Description	General		
Digital Video	Multimedia	2nd	
Event Core	General Event Processing		
Event Processing	Event Processing General	3rd	refers to the IoT and sensors domain
Gear Species		4th	refers to the fishery domain
Gear Vessel	Fishery	2nd	
Gear Water Area	Fishery	2nd	
GO Top	Biology	2nd	
Hazardous Situation	General Event Processing Participation	3rd	the *hazardous situation* concept does not apply to the data protection domain
Information Realization	Semiotics		
Intension Extension	General Semiotics		
Invoice	Business	2nd	
List			
Literal Reification		3rd	more focused on solving expressivity problems in OWL

(Continue on the next page)

CP Name	Domain	It.	Explanation
Map Legend Ontology	Geography GIScience	2nd	
Nary Participation	General		
News Reporting Event	Event Processing Media Social Science	3rd	too specific for being applicable in the domain of interest
Object with states	General		
Object Role	General		
Observation	General Science	3rd	refers to scientific experiments
Parameter	General		
Part Of	Parts and Collections		
Participant Role	General Organization		
Participation	General		
Periodic Interval	Time		
Place	General		
Policy	General		
Privacy Policy Personal Data			
Reaction	Workflow		
Recurrent Event Series			
Recurrent Situation Series	General		
Region			
Reporting Event	General Event Processing	3rd	refers to contradictory descriptions of an event
Reporting News Event	Event Processing Media Social Science	3rd	refers to contradictory descriptions of an event
Resource Abundance Observation	Fishery	2nd	
Resource Exploitation Observation	Fishery	2nd	
Role task	Organization Management Scheduling		
Sequence	General Organization Workflow Time		
Set			
Simple Or Aggregated	Parts and Collections	3rd	could be related to simpler patterns, as also noticed by the reviewers of the pattern
Situation	General		

(Continue on the next page)

CP Name	Domain	It.	Explanation
Smart Home Feature Of Interest		4th	refers to the IoT and sensors domain
Smart Home Geometry		4th	refers to the IoT and sensors domain
Smart Home Network		4th	refers to the IoT and sensors domain
Smart Home Object		4th	refers to the IoT and sensors domain
Smart Home Place		4th	refers to the IoT and sensors domain
Smart Home Property		4th	refers to the IoT and sensors domain
Smart Home Sensing		4th	refers to the IoT and sensors domain
Smart Home Situation		4th	out of scope
Smart Home Time Interval		4th	out of scope
Spatio Temporal Extent	Earth Science or Geoscience General	2nd	
Species Bathymetry		4th	refers to the natural science domain
Species Conditions		4th	refers to the natural science domain
Species Conservation		4th	refers to the natural science domain
Species Eat		4th	refers to the natural science domain
Species Habitat		4th	refers to the natural science domain
Species Names		4th	refers to the natural science domain
Standard Enforcer Pattern		4th	refers to compliance with standards, too specific for the domain of interest
Tagging	General Web2.0 Document Management	2nd	
Task Execution	Organization Management Scheduling Workflow		
Time indexed participation	General		
Time Indexed Part Of	Parts and Collections		
Time Indexed Situation	General		
Time Interval	Time		
Time Period	Time		
Topic	General		
Trajectory	General Earth Science or Geoscience	2nd	

(Continue on the next page)

CP Name	Domain	It.	Explanation
Transition	General Workflow Manufacturing	2^{nd}	
Types of entities	General		
Vertical Distribution		4^{th}	refers to the fishery domain
Vessel Species	Fishery	2^{nd}	
Vessel Water Area		4^{th}	refers to the fishery domain

B Annotation Guidelines

B.1 Introduction

Dear Annotator, thank you very much for accepting my request to participate in this annotation task. You will be asked to read a set of sentences extracted from privacy policies and answer some questions related to the information provided by those sentences. This document will provide you with some background knowledge to put the task in context and with the guidelines that you are required to follow in your task.

B.2 Background

B.2.1 The General Data Protection Regulation (GDPR)

The General Data Protection Regulation (GDPR) is the EU regulation that set consistent data protection rules across Europe. It applies to all companies that process personal data about individuals in the EU. You can find the full text of the Regulation in the Eur-Lex portal[1]. This section recalls some of the main aspects of the Regulation. Specifically, Article 4 provides, among the others, the following definitions:

- *personal data* means any information relating to an identified or identifiable natural person ("data subject"); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier 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;
- *processing* means any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction.
- *controller* means the natural or legal person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data; where the purposes and means of such processing are determined by Union or Member State law, the controller or the specific criteria for its nomination may be provided for by Union or Member State law;

[1] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A02016R0679-20160504

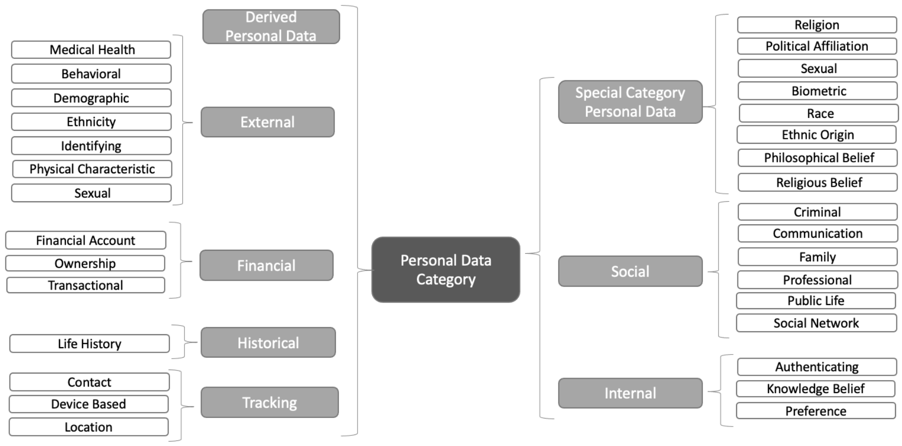


Figure B.1: The hierarchical structure of the *Personal Data Category* module. The Figure only shows the first levels of the hierarchy, while the DPV further specialises the concepts in the white boxes.

- *processor* means a natural or legal person, public authority, agency or other body which processes personal data on behalf of the controller;
- *recipient* means a natural or legal person, public authority, agency or another body, to which the personal data are disclosed, whether a third party or not. However, public authorities [...] in the framework of a particular inquiry in accordance with Union or Member State law shall not be regarded as recipients; the processing of those data [...] shall be in compliance with the applicable data protection rules according to the purposes of the processing;
- *third party* means a natural or legal person, public authority, agency or body other than the data subject, controller, processor and persons who, under the direct authority of the controller or processor, are authorised to process personal data;

The processing of personal data should be undertaken lawfully, fairly and transparently with respect to the data subject, without diverging from the original purpose of the processing (purpose limitation).

In order for the processing of personal data to be lawful, it must be based on six main *legal grounds*, introduced in Article 6: the data subject provided her consent for the processing; the processing is necessary to enter into or perform a contract with the data subject; the processing is necessary to protect the vital interest of some individuals or to comply with a legal obligation; the processing is necessary for public interest under EU or national law; the processing is necessary for the legitimate interest of the controller or the third party, when the processing doesn't impact on the fundamental rights and freedoms of the data subject.

According to Article 13 and Article 14, the data subject must be provided with specific information when her personal data are processed. Some of this information concerns the categories of personal data being processed, the purpose of the processing, the legal ground for processing and the possible recipients of the personal data.

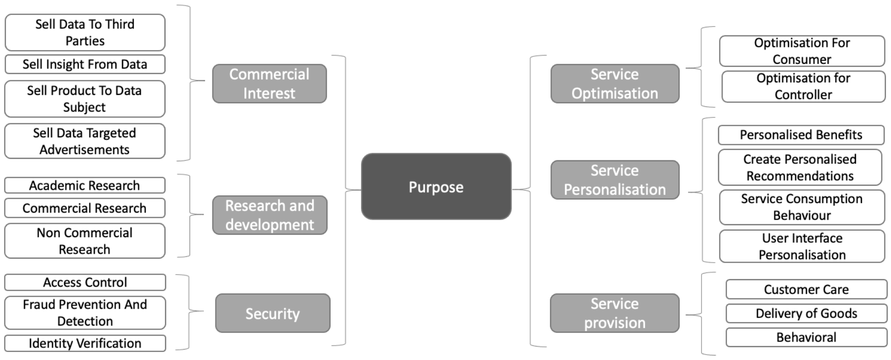


Figure B.2: The hierarchical structure of the *Purpose* module. The Figure only shows the first levels of the hierarchy, while the DPV further specialises the concepts in the white boxes.

B.2.2 The Data Privacy Vocabulary (DPV)

The Data Privacy Vocabulary (DPV) is an on-line resource[2] that collects and organises the concepts related to the different aspects involved in the processing of personal data, with a reference to the regulatory framework set by the GDPR. It is organised in several modules, each of them referring to a specific aspect concerning the Regulation. You should focus on three modules: *Personal Data Category* that represents different categories of personal data, *Purpose* that represents the reason why the personal data are processed and *Processing* that represents different types of processing that can be performed on data.

In each module, the concepts are organised in a *hierarchy*, i.e. they are linked by generalisation/specialisation relations. Figure B.1, Figure B.2 and Figure B.3 show visual representations of the organisation of concepts within the modules. In each figure, the concepts surrounded by a dark-grey shaded box are the most generic concepts of the modules (and their names correspond to the modules names). Concepts that are surrounded by a light-grey shaded box specialise the meaning of the most generic concepts, while the concepts in white boxes further specialise the corresponding concepts in the light grey-shaded boxes. For instance, in Figure B.1 the *Location* concept (in the bottom left corner), specialises the meaning of the *Traking* concept that, in turn, specialises the meaning of the most general concept *Personal Data*. Of course, given this hierarchical organisation, we can also intuitively infer that *Location* is a concept more specific than *Personal Data Category*. Please, note that Figure B.1 and Figure B.2 do not show all the concepts in the corresponding modules. This means that, the concepts in the withe boxes are further specialised by other concepts in the DPV, but they were omitted in the figures to ease the visualisation. By contrast, Figure B.3 shows the entire hierarchy of concepts in the *Processing* module.

Each concept in the DPV is associated to a short description of the meaning that it represents. For instance, the *Tracking* concept is described as *"Personal data that can be used to track an individual or used as an identifier, e.g. location or email"*. Each concept is also associated to and URI. Having the URI of a concept, you can use it to visualise its description and navigate the hierarchical structure in which it is inserted. The URI of each concept in the DPV starts with "http://www.w3.org/ns/dpv#", followed by the name of the concept. For instance, the concept *Location* is associated with the URI `http://www.w3.org/ns/dpv#Location`. By clicking on this address, you will be redirected to the Web page of the DPV, in the specific section where the concept

[2] `https://dpvcg.github.io/dpv/`

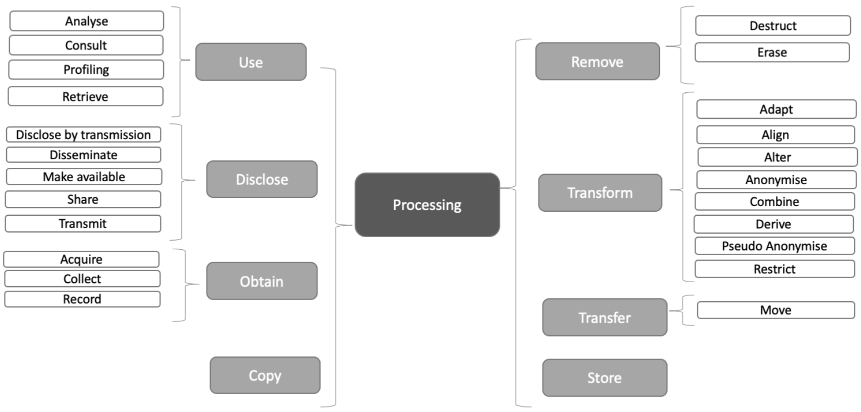
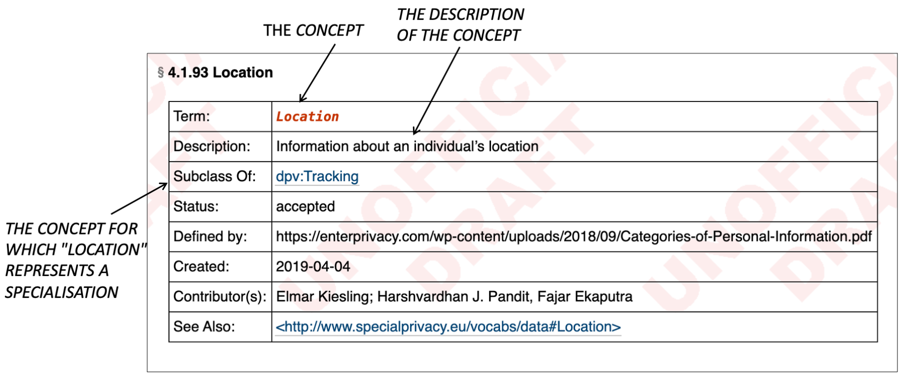


Figure B.3: The hierarchical structure of the *Processing* module. The Figure shows all the concepts of the module.



§ 4.1.93 Location

Term:	Location
Description:	Information about an individual's location
Subclass Of:	dpv:Tracking
Status:	accepted
Defined by:	https://enterprivacy.com/wp-content/uploads/2018/09/Categories-of-Personal-Information.pdf
Created:	2019-04-04
Contributor(s):	Elmar Kiesling; Harshvardhan J. Pandit, Fajar Ekaputra
See Also:	<http://www.specialprivacy.eu/vocabs/data#Location>

Figure B.4: A section of the DPV Web page, dedicated to the description of the *Location* concept.

is defined. Figure B.4 shows the section in the DPV Web page dedicated to the *Location* concept.

B.3 The annotation task

B.3.1 An introduction to the system

The annotations that you will provide by performing the task will be used to evaluate the performance of a system that automatically identifies certain information expressed by privacy policies on which the legal framework set by the GDPR applies.

Specifically, for a certain sentence of a privacy policy, the system tries to identify the type of processing the sentence describes[3]. Then, according to the type of processing that is identified, the

[3]When I mention the term *processing*, I refer to the meaning that the term assumes according to the definition given by the GDPR, reported in the Section B.2.1

system extracts some more detailed information about the processing activity.

The information that the system identifies is collected inside a *processing template*. Three types of processing template can be associated to a sentence, according to the processing activitiy it describes: *obtain*, *disclose* and *other processing*. In each template, the information is organised in *template components* that specify the details about the processing activity emerging from the sentence.

Below, you can find a description of the processing activities that each template aims to represent, and the template components that characterise it.

***Obtain* template.**

This template applies when the sentence describes those processing activities aimed at gathering personal data of individuals that use the services/goods offered by the company. The components that characterise this template are:

- **processing type**: specifies the verb that, inside the sentence, expresses the activity through which the company gathers the personal data of its users;
- **personal data**: refers to the personal data on which the processing is performed;
- **obtains data from agent**: specifies the party from which the personal data are obtained;
- **purpose**: refers to the purpose for which the personal data are processed.
- **mechanism for obtaining data**: specifies whether the personal data are obtained by an automated mean that does not ask for human involvement.

***Disclose* template.**

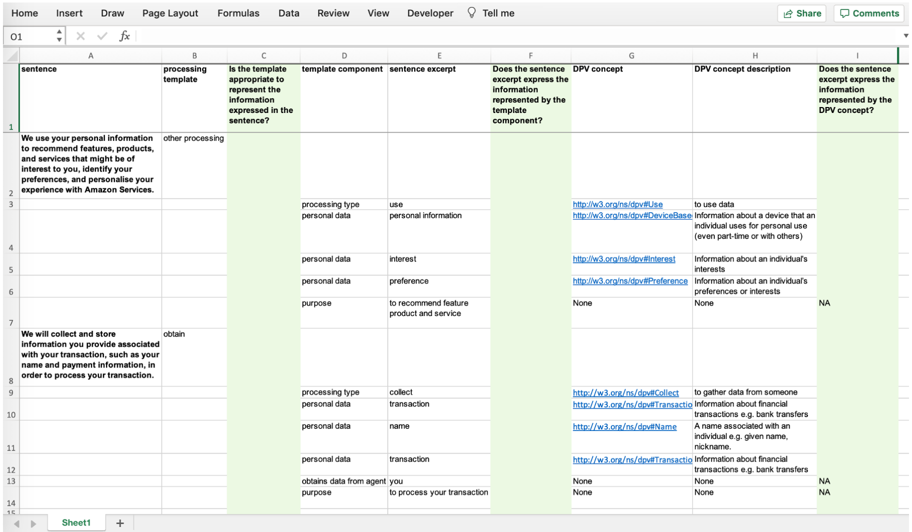
This template applies when the sentence describes the processing activities related to the disclosure of personal data of individuals to parties other than the data subject (e.g., other companies or organisations). The components that characterise this template are:

- **processing type**: specifies the verb that, inside the sentence, expresses the activity through which the company discloses personal data;
- **personal data**: refers to the personal data on which the processing is performed;
- **discloses data to agent**: specifies the recipients of the personal data;
- **purpose**: refers to the purpose for which the personal data are processed.

***Other processing* template.**

This template applies when the sentence concerns the processing activities other than those represented by the previous templates. The components that characterise this template are:

- **processing type**: specifies the verb that, inside the sentence, expresses the processing activity that the company performs;
- **personal data**: refers to the personal data on which the processing is performed;
- **purpose**: refers to the purpose for which the personal data are processed.



	sentence	processing template	Is the template appropriate to represent the information expressed in the sentence?	template component	sentence excerpt	Does the sentence excerpt express the information represented by the template component?	DPV concept	DPV concept description	Does the sentence excerpt express the information represented by the DPV concept?
2	We use your personal information to recommend features, products, and services that might be of interest to you, identify your preferences, and personalise your experience with Amazon Services.	other processing							
3				processing type	use		http://w3.org/ns/dpv#Use	to use data	
4				personal data	personal information		http://w3.org/ns/dpv#DeviceBase	Information about a device that an individual uses for personal use (even part-time or with others)	
5				personal data	interest		http://w3.org/ns/dpv#Interest	Information about an individual's interests	
6				personal data	preference		http://w3.org/ns/dpv#Preference	Information about an individual's preferences or interests	
7				purpose	to recommend feature product and service		None	None	NA
8	We will collect and store information you provide associated with your transaction, such as your name and payment information, in order to process your transaction.	obtain							
9				processing type	collect		http://w3.org/ns/dpv#Collect	to gather data from someone	
10				personal data	transaction		http://w3.org/ns/dpv#Transactio	Information about financial transactions e.g. bank transfers	
11				personal data	name		http://w3.org/ns/dpv#Name	A name associated with an individual e.g. given name, nickname.	
12				personal data	transaction		http://w3.org/ns/dpv#Transactio	Information about financial transactions e.g. bank transfers	
13				obtains data from agent	you		None	None	NA
14				purpose	to process your transaction		None	None	NA

Figure B.5: The file that will contain your annotation.

As you probably noticed, some template elements are common to all the processing templates (i.e. *processing type, personal data* and *purpose*), while other template elements are specific of a template. A sentence could express multiple information and, consequently, fit one or more template. For helping you to understand how the processing templates could be used to organise the information expressed by a privacy policy, consider the following sentence, extracted from a privacy policy:

> *We will collect and store information you provide associated with your transaction, such as your name and payment information, in order to process your transaction.*

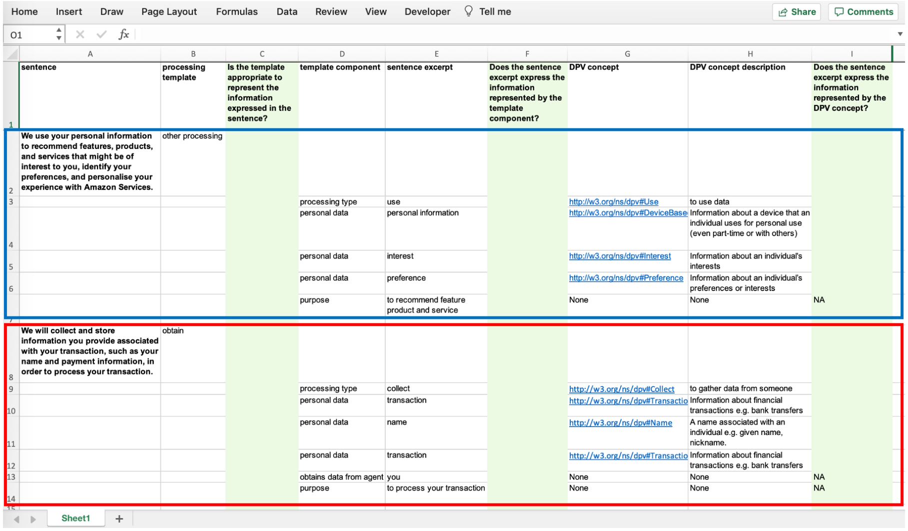
We expect two templates to apply to this sentence: the *obtain* template and the *other processing* template. The *obtain* template is applicable because the sentence refers to the collection of personal data. The verb that specifies this processing type is *collect*. The sentence also mentions that the personal data being collected are those *associated to a transaction* of the user, including *name* and *payment information*. The data are collected from the user (*"information you provide"*) and they are collected for the purpose of processing the transaction (*"in order to process your transction"*). In the meantime, the *other processing* template could be also applicable to the sentence, with reference to the processing activity of storing personal data. The verb that specifies this processing activity is *store*. The personal data being collected, the agent that provides those data and the purpose of the processing are the same identified for the previous template.

Sometimes a single sentence does not express the information that is necessary to fill all the components of a template. For instance, the purpose of the processing or the parties to which personal data are disclosed could miss. Consequently, a processing template applied to a sentence could specify only a subset of the components that characterise it.

For the *processing type*, *personal data* and *purpose* components of the three processing templates, the system also tries to associate a corresponding concept in the DPV. Considering the sentence above, the *collect* and the *store* verbs could be mapped on the corresponding concepts *Collect*[4] and *Store*[5] in the DPV. Similarly, the personal data that correspond to the name of the

[4] http://www.w3.org/ns/dpv#Collect

[5] http://www.w3.org/ns/dpv#Store



	A	B	C	D	E	F	G	H	I
1	**sentence**	**processing template**	**Is the template appropriate to represent the information expressed in the sentence?**	**template component**	**sentence excerpt**	**Does the sentence excerpt express the information represented by the template component?**	**DPV concept**	**DPV concept description**	**Does the sentence excerpt express the information represented by the DPV concept?**
2	**We use your personal information to recommend features, products, and services that might be of interest to you, identify your preferences, and personalise your experience with Amazon Services.**	other processing							
3				processing type	use		http://w3.org/ns/dpv#Use	to use data	
4				personal data	personal information		http://w3.org/ns/dpv#DeviceBase	Information about a device that an individual uses for personal use (even part-time or with others)	
5				personal data	interest		http://w3.org/ns/dpv#Interest	Information about an individual's interests	
6				personal data	preference		http://w3.org/ns/dpv#Preference	Information about an individual's preferences or interests	
7				purpose	to recommend feature product and service		None	None	NA
8	**We will collect and store information you provide associated with your transaction, such as your name and payment information, in order to process your transaction.**	obtain							
9				processing type	collect		http://w3.org/ns/dpv#Collect	to gather data from someone	
10				personal data	transaction		http://w3.org/ns/dpv#Transactio	Information about financial transactions e.g. bank transfers	
11				personal data	name		http://w3.org/ns/dpv#Name	A name associated with an individual e.g. given name, nickname.	
12				personal data	transaction		http://w3.org/ns/dpv#Transactio	Information about financial transactions e.g. bank transfers	
13				obtains data from agent	you		None	None	NA
14				purpose	to process your transaction		None	None	NA

Figure B.6: The blue box and the red box represent two different processing templates in the file Excel.

data subject could be represented by the *Name*[6] concept in the DPV. The association of a template element to a concept in the DPV is not always specified. Some elements in a template may be associated with a concept in the DPV, while others may not.

B.3.2 Your task

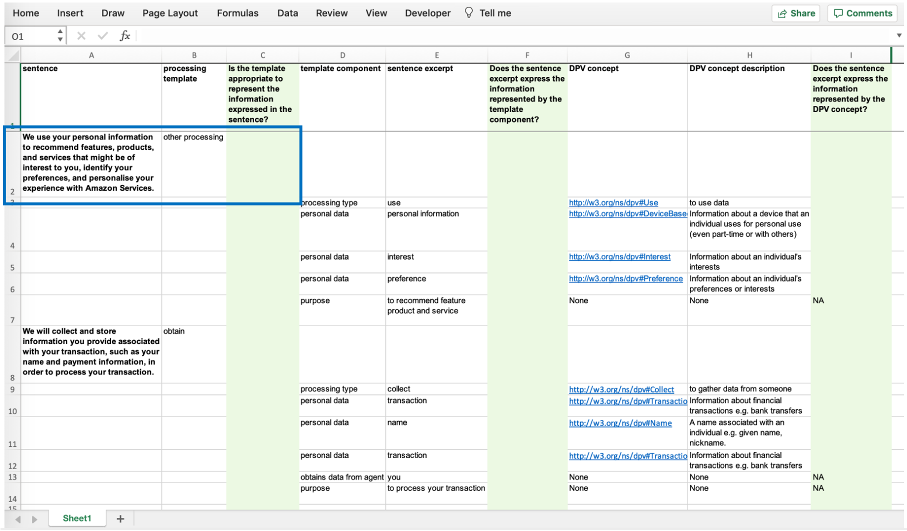
You will be provided with a file having the *.xlsx* extension, that you can easily open using the Microsoft Excel software. Once the file is open, its appearance should be similar to the one shown in Figure B.5. The file is made of several columns and three of them are highlighted in light green to indicate the columns where you will be asked to insert your annotation.

The first row of the file contains the headers of the columns to guide you in understanding the information that have been inserted in the file. The file contains the processing templates that have been extracted by the system from a set of privacy policies. You can visually identify the type templates looking at column A in Figure B.5. Every time that a cell in column A contains a sentence, a new processing template begins. Consequently, all the subsequent rows that are blank in column A indicate that the information contained in columns from B to I refer to the same sentence in column A. Figure B.6 shows how the processing templates are organised in the file grid.

Below, the steps necessary to carry out your annotation task will be explained. For each processing template, you should follow a three-step flow that will guide you in answering three questions.

Step 1) Evaluation of the overall suitability of a processing template. For a certain processing template, in the first step, you should focus on columns A, B and C, as highlighted in Figure B.7.

[6] http://www.w3.org/ns/dpv#Name



sentence	processing template	Is the template appropriate to represent the information expressed in the sentence?	template component	sentence excerpt	Does the sentence excerpt express the information represented by the template component?	DPV concept	DPV concept description	Does the sentence excerpt express the information represented by the DPV concept?
We use your personal information to recommend features, products, and services that might be of interest to you, identify your preferences, and personalise your experience with Amazon Services.	other processing							
			processing type	use		http://w3.org/ns/dpv#Use	to use data	
			personal data	personal information		http://w3.org/ns/dpv#DeviceBase	Information about a device that an individual uses for personal use (even part-time or with others)	
			personal data	interest		http://w3.org/ns/dpv#Interest	Information about an individual's interests	
			personal data	preference		http://w3.org/ns/dpv#Preference	Information about an individual's preferences or interests	
			purpose	to recommend feature product and service		None	None	NA
We will collect and store information you provide associated with your transaction, such as your name and payment information, in order to process your transaction.	obtain							
			processing type	collect		http://w3.org/ns/dpv#Collect	to gather data from someone	
			personal data	transaction		http://w3.org/ns/dpv#Transactio	Information about financial transactions e.g. bank transfers	
			personal data	name		http://w3.org/ns/dpv#Name	A name associated with an individual e.g. given name, nickname.	
			personal data	transaction		http://w3.org/ns/dpv#Transactio	Information about financial transactions e.g. bank transfers	
			obtains data from agent	you		None	None	NA
			purpose	to process your transaction		None	None	NA

Figure B.7: Columns you should focus on in Step 1 of your annotation task.

Read carefully the sentence in column A. Then, move to the next column on the right (i.e. column B) and read the processing template that has been assigned to the sentence. For instance, in Figure B.7, the sentence in cell A2 explains how the personal information of a user are processed to personalise the services offered by the company according to her preferences and interests. This sentence is assigned to the processing template called *other processing*.

On the same row where you read the sentence and the processing template label, moving to the right, you will find the first column highlighted in green (i.e. column C). Here you are required to provide an answer to the following question:

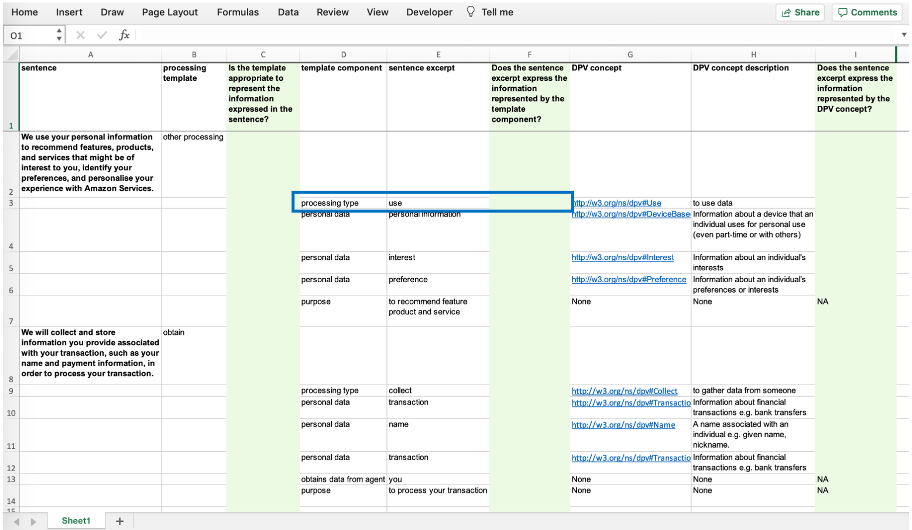
Is the template appropriate to represent the information expressed in the sentence?

For answering the question, you can in any moment consult the description of the processing templates provided in Section B.3.1. The question admits only two types of answer:

- type YES, if you think that the processing template is appropriate with respect to the information provided by the sentence;
- type NO, if you think that the processing template is not appropriate with respect to the information provided by the sentence.

Please, insert your answer in column C, on the same row where you read the sentence. In Figure B.7, for instance, the answer should be provided in cell C2. An answer must always be specified and blank answers are not permitted. As anticipated in Section B.3.1, a sentence could fit more that one processing template. If this is the case, you will find a distinct processing template for each processing activity identified by the system. However, your answer to the question should be solely based on the information specified in the specific processing template that you are considering, avoiding looking to the other processing templates.

Sometimes sentences could explicitly refer to processing activities that are not performed, e.g. a sentence might state *we do not collect information related to your credit card*. In this case, the association of the *obtain* template to the sentence should be considered appropriate because the sentence describes a collection of personal data, even if such a processing activity is not carried out by the company.



	A	B	C	D	E	F	G	H	I
1	sentence	processing template	Is the template appropriate to represent the information expressed in the sentence?	template component	sentence excerpt	Does the sentence excerpt express the information represented by the template component?	DPV concept	DPV concept description	Does the sentence excerpt express the information represented by the DPV concept?
2	We use your personal information to recommend features, products, and services that might be of interest to you, identify your preferences, and personalise your experience with Amazon Services.	other processing							
3				processing type	use		ttp://w3.org/ns/dpv#Use	to use data	
4				personal data	personal information		http://w3.org/ns/dpv#DeviceBase	Information about a device that an individual uses for personal use (even part-time or with others)	
5				personal data	interest		http://w3.org/ns/dpv#Interest	Information about an individual's interests	
6				personal data	preference		http://w3.org/ns/dpv#Preference	Information about an individual's preferences or interests	
7				purpose	to recommend feature product and service		None	None	NA
8	We will collect and store information you provide associated with your transaction, such as your name and payment information, in order to process your transaction.	obtain							
9				processing type	collect		http://w3.org/ns/dpv#Collect	to gather data from someone	
10				personal data	transaction		http://w3.org/ns/dpv#Transactio	Information about financial transactions e.g. bank transfers	
11				personal data	name		http://w3.org/ns/dpv#Name	A name associated with an individual e.g. given name, nickname.	
12				personal data	transaction		http://w3.org/ns/dpv#Transactio	Information about financial transactions e.g. bank transfers	
13				obtains data from agent	you		None	None	NA
14				purpose	to process your transaction		None	None	NA

Figure B.8: Columns to focus on in Step 2 of your annotation task.

Step 2) Evaluate a template component. In the next step, you should focus on columns D, E and F, as shown in Figure B.8. Read carefully, on column D, a template component that refers to the processing template that you read in Step 1. Then, on the same row, in column E, read the sentence excerpt associated to the template component, remembering that the sentence excerpt is a piece of the sentence that you read in Step 1, on column A. Then, answer the following question:

> *Does the sentence excerpt express the information represented by the template component?*

Also in this step, you are encouraged to consult the descriptions of the templates components provided for each template in Section B.3.1. The question admits only two types of answer:

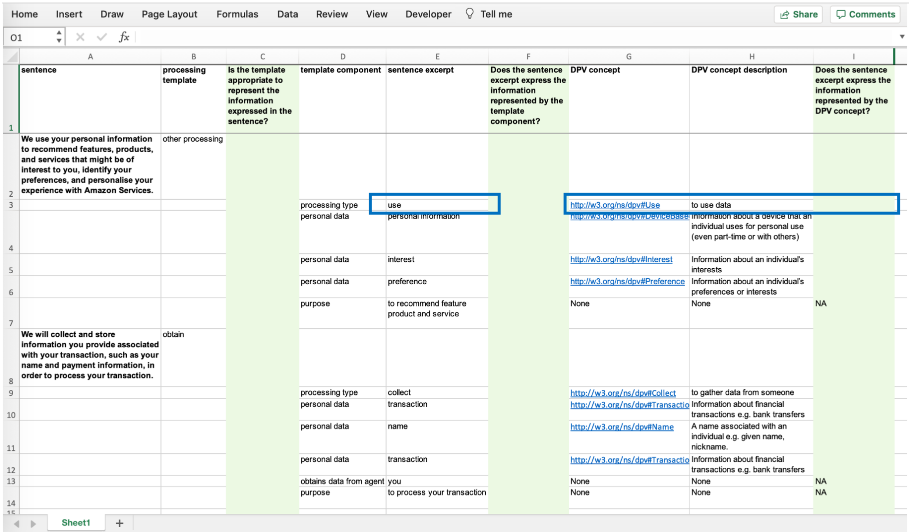
- type YES, if you think that the sentence excerpt is appropriate to represent the information expressed by the template component. A sentence excerpt may be deemed appropriate when it expresses only part of the complete information that the fragment contains[7], or when it only provides a shallow mention the information that the template component represents. A sentence excerpt may also be deemed appropriate when it expresses more information than it is needed in that template component.

- type NO, if you think that the sentence excerpt does not represent the information expressed by the template component.

Please, insert your answer in column F, on the same row where you read the template component and the corresponding sentence excerpt. An answer must always be specified and blank answers are not permitted.

If you found, in Step 1, that the processing template associated to the sentence is not appropriate for the sentence (and, consequently, you answered NO to the question in Step 1), then you

[7] In the sentence *"we collect your personal data to personalise your user experience and to send you marketing communications"*, the sentence excerpt *"to personalise your user experience"* expresses the information that the *purpose* element template represents, even if the sentence also specify another purpose (*"to send you marketing communications"*).



	A	B	C	D	E	F	G	H	I
1	**sentence**	**processing template**	**Is the template appropriate to represent the information expressed in the sentence?**	**template component**	**sentence excerpt**	**Does the sentence excerpt express the information represented by the template component?**	**DPV concept**	**DPV concept description**	**Does the sentence excerpt express the information represented by the DPV concept?**
2	**We use your personal information to recommend features, products, and services that might be of interest to you, identify your preferences, and personalise your experience with Amazon Services.**	other processing							
3				processing type	use		http://w3.org/ns/dpv#Use	to use data	
4				personal data	personal information		http://w3.org/ns/dpv#DeviceBase	Information about a device that an individual uses for personal use (even part-time or with others)	
5				personal data	interest		http://w3.org/ns/dpv#Interest	Information about an individual's interests	
6				personal data	preference		http://w3.org/ns/dpv#Preference	Information about an individual's preferences or interests	
7				purpose	to recommend feature product and service		None	None	NA
8	**We will collect and store information you provide associated with your transaction, such as your name and payment information, in order to process your transaction.**	obtain							
9				processing type	collect		http://w3.org/ns/dpv#Collect	to gather data from someone	
10				personal data	transaction		http://w3.org/ns/dpv#Transactio	Information about financial transactions e.g. bank transfers	
11				personal data	name		http://w3.org/ns/dpv#Name	A name associated with an individual e.g. given name, nickname.	
12				personal data	transaction		http://w3.org/ns/dpv#Transactio	Information about financial transactions e.g. bank transfers	
13				obtains data from agent	you		None	None	NA
14				purpose	to process your transaction		None	None	NA

Figure B.9: Columns to focus on in Step 3 of your annotation task

can type NA, to indicate that the answer is *not applicable* in that case.

Sometimes, you could notice some differences between the sentence excerpt and the whole sentence. This happens because the text in the sentence excerpt is lemmatised, i.e. it is written in its standard grammatical form (e.g. verbs are in their infinitive forms, nouns are in their singular form, etc.). For example, the phrase *"our business partners"* is lemmatised to *"we business partner"*. These differences should not be considered in your assessment. For each excerpt, even if lemmatised, you should be able to identify its corresponding part in the sentence.

Step 3) Evaluate the association with a concept in the DPV. In the last step, you should focus on columns E, G, H and I, as shown in Figure B.9. Consider again the sentence excerpt that you read on column E in Step 2. Then, read on column G the name of the concept in the DPV associated to the sentence excerpt and read in column H the description of the concept provided by the DPV[8]. Then, answer the following question:

> *Does the sentence excerpt express the information represented by the DPV concept?*

The question admits only two types of answer:

- type YES, if you think that the sentence excerpt adheres to the description of the concept in the DPV;
- type NO, if you think that the sentence excerpt does not corresponds to the description of the concept in the DPV.

Please, insert your answer in column I, on the same row where you read the DPV concept and its description. An answer must always be specified and blank answers are not permitted.

[8]The description is the same that you would read by clicking on the URI of the concept. The description appears on the annotation file to make your work easier, so that you do not have to switch from the annotation file to the browser. However, you can click on the URI and access the DPV page if you need to have more information about the concept, for instance, by browsing the concepts hierarchy in which it is placed.

As mentioned in Section B.3.1, not all the template components and the corresponding sentence excerpts are associated to a DPV concept. If this is the case, you will find a pre-define value *None* to indicate the absence of such a mapping. Moreover, the pre-defined value NA (i.e. not applicable) is inserted in column I, to indicate that an answer to this question is not requested. You can find an example of this situation in Figure B.9, looking at row 7.

If you found, in Step 1, that the processing template associated to the sentence is not appropriate for the sentence (and, consequently, you answered NO to the question in Step 1) or if you found, in Step 2, that the sentence excerpt does not properly express the information of the template component, then you can type NA in column I, to indicate that the answer is *not applicable* in that case.

How to continue. Step 2 and Step 3 should be repeated for each template component specified for a processing template. Sometimes you could find, in Step 2, the same template component associated to the same text excerpt multiple times, as it is the case in the second processing template that was highlighted in Figure B.6 (see rows 10 and 12 in the Figure). This happens when a sentence has multiple occurrences of the same term. Please, provide your answer to each occurrence and continue your evaluation also for Step 3.

When you have answered all the questions for all the parts of a processing template, you can start annotating a new template, following the steps from 1 to 3. Please, remember that, in any step of the annotation, the answers to your questions should be provided *independently* from the information that you could find in other processing templates or in other templates components. Proceed one row at a time, as explained in these guidelines and represented in Figures B.7, B.8, B.9.

Together with the file to be annotated, you are provided also with the file that was used to show the annotation process in the figures of this guide, filled with possible answers to show you the expected outcome of your work.

Overall, the annotation file contains 75 sentences that has been associated to as many type templates. When sending back the results, I would also kindly ask to indicate:

- the time it took you to read this guidelines;
- the time that it took you to annotate, on average, a processing template when you annotated the first ten processing templates in the file;
- the time that it took you to annotate, on average, a processing template when you annotated the following ten processing templates in the file;
- the time that it took you to annotate, on average, a processing template when you annotated the remaining processing templates in the file.

Thank you again for your collaboration. Good work!

Bibliography

[1] 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 L119. p. 1-88 (May 2016).

[2] A. S. Abdelghany, N. R. Darwish, and H. A. Hefni. An agile methodology for ontology development. *International Journal of Intelligent Engineering and Systems*, 12(2):170–181, 2019.

[3] G. Aguado de Cea, A. Gómez-Pérez, E. Montiel-Ponsoda, and M. C. Suárez-Figueroa. Natural language-based approach for helping in the reuse of ontology design patterns. In A. Gangemi and J. Euzenat, editors, *Knowledge Engineering: Practice and Patterns*, pages 32–47, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg.

[4] G. Ajani, G. Boella, L. Di Caro, L. Robaldo, L. Humphreys, S. Praduroux, P. Rossi, and A. Violato. The european legal taxonomy syllabus: a multilingual, multi-level ontology framework to untangle the web of european legal terminology. *Applied Ontology*, 11(4):325–375, 2016.

[5] A. Akbik and A. Löser. KrakeN: N-ary facts in open information extraction. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction (AKBC-WEKEX)*, pages 52–56, Montréal, Canada, jun 2012. Association for Computational Linguistics.

[6] R. Amos, G. Acar, E. Lucherini, M. Kshirsagar, A. Narayanan, and J. Mayer. Privacy policies over time: Curation andanalysis of a million-document dataset. 2020.

[7] A. I. Anton, J. B. Earp, Q. He, W. Stufflebeam, D. Bolchini, and C. Jensen. Financial privacy policies and the need for standardization. *IEEE Security & privacy*, 2(2):36–45, 2004.

[8] G. Antoniou and F. Van Harmelen. *A Semantic Web Primer*, chapter 1, pages 1–24. John Wiley & Sons, second edition, 2008.

[9] R. Artstein and M. Poesio. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596, 2008.

[10] M. N. Asim, M. Wasim, M. U. G. Khan, W. Mahmood, and H. M. Abbasi. A survey of ontology learning techniques and applications. *Database*, 2018, 10 2018. bay101.

[11] Assembly Bill No. 375 Chapter 55: An act to add Title 1.81.5 (commencing with Section 1798.100) to Part 4 of Division 3 of the Civil Code, relating to privacy. California State Legislature, Jun. 29, 2018, Key: Assembly Bill No. 375. [Online]. https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201720180AB375. Accessed: 2020-11-16.

[12] T. Athan, G. Governatori, M. Palmirani, A. Paschke, and A. Wyner. Legalruleml: Design principles and foundations. In *Reasoning Web International Summer School*, pages 151–188. Springer, 2015.

[13] C. F. Baker, C. J. Fillmore, and J. B. Lowe. The berkeley framenet project. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 86–90, 1998.

[14] M. Banko, M. J. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni. Open information extraction from the web. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, IJCAI'07, page 2670–2676, San Francisco, CA, USA, 2007. Morgan Kaufmann Publishers Inc.

[15] C. Bartolini, A. Giurgiu, G. Lenzini, and L. Robaldo. Towards legal compliance by correlating standards and laws with a semi-automated methodology. In *Benelux Conference on Artificial Intelligence*, pages 47–62. Springer, 2016.

[16] T. Berners-Lee, J. Hendler, and O. Lassila. The semantic web. *Scientific american*, 284(5):34–43, 2001.

[17] A. Bernstein, J. Hendler, and N. Noy. A new look at the semantic web. *Communications of the ACM*, 59(9):35–37, 2016.

[18] J. Bhatia and T. D. Breaux. Semantic incompleteness in privacy policy goals. In *2018 IEEE 26th International Requirements Engineering Conference (RE)*, pages 159–169. IEEE, 2018.

[19] J. Bhatia, T. D. Breaux, J. R. Reidenberg, and T. B. Norton. A theory of vagueness and privacy risk perception. In *2016 IEEE 24th International Requirements Engineering Conference (RE)*, pages 26–35. IEEE, 2016.

[20] J. Bhatia, M. C. Evans, and T. D. Breaux. Identifying incompleteness in privacy policy goals using semantic frames. *Requirements Engineering*, 24(3):291–313, 2019.

[21] A. Björkelund, L. Hafdell, and P. Nugues. Multilingual semantic role labeling. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL 2009): Shared Task*, pages 43–48, 2009.

[22] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[23] E. Blomqvist, V. Presutti, E. Daga, and A. Gangemi. Experimenting with extreme design. In *International Conference on Knowledge Engineering and Knowledge Management*, pages 120–134. Springer, 2010.

[24] G. Boella and T. v. d. L. Regulative. Constitutive norms in normative multiagent systems. In *KR'04 Proceedings of the Ninth International Conference on Principles of Knowledge Representation and Reasoning*, pages 255–265.

[25] P. A. Bonatti, S. Kirrane, I. M. Petrova, and L. Sauro. Machine understandable policies and gdpr compliance checking. *[online]*, 2020.

[26] J. Brank, M. Grobelnik, and D. Mladenic. A survey of ontology evaluation techniques. In *Proceedings of the conference on data mining and data warehouses (SiKDD 2005)*, pages 166–170. Citeseer Ljubljana, Slovenia, 2005.

[27] J. Breuker, R. Hoekstra, et al. Core concepts of law: taking commonsense seriously. In *Proceedings of formal ontologies in information systems (FOIS-2004)*, pages 210–221, 2004.

[28] M. G. Buey, A. L. Garrido, C. Bobed, and S. Ilarri. The ais project: Boosting information extraction from legal documents by using ontologies. In *ICAART (2)*, pages 438–445, 2016.

[29] M. Burri and R. Schär. The reform of the EU data protection framework: outlining key changes and assessing their fitness for a data-driven economy. *Journal of Information Policy*, 6(1):479–511, 2016.

[30] A. Burton-Jones, V. C. Storey, V. Sugumaran, and P. Ahluwalia. A semiotic metrics suite for assessing the quality of ontologies. *Data & Knowledge Engineering*, 55(1):84–102, 2005.

[31] V. A. Carriero, A. Gangemi, M. L. Mancinelli, A. G. Nuzzolese, V. Presutti, and C. Veninata. Pattern-based design applied to cultural heritage knowledge graphs. `https://arxiv.org/abs/1911.07585#cs.AI`, 2020.

[32] P. Casanovas, M. Palmirani, S. Peroni, T. Van Engers, and F. Vitali. Semantic web for the legal domain: the next step. *Semantic Web*, 7(3):213–227, 2016.

[33] N. Casellas. *Legal ontology engineering: Methodologies, modelling trends, and the ontology of professional judicial knowledge*, volume 3. Springer Science & Business Media, 2011.

[34] I. Chalkidis. Law2vec: Legal word embeddings. `https://archive.org/details/Law2Vec`, 2018. Accessed: 2020-11-30.

[35] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos. Legal-bert: The muppets straight out of law school, 2020.

[36] D. V. Cicchetti and S. A. Sparrow. Developing criteria for establishing interrater reliability of specific items: applications to assessment of adaptive behavior. *American journal of mental deficiency*, 1981.

[37] P. Cimiano. *Ontology Learning and Population from Text: Algorithms, Evaluation and Applications*. Springer, 2006.

[38] P. Cimiano, A. Mädche, S. Staab, and J. Völker. *Ontology Learning*, pages 245–267. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009.

[39] P. Cimiano and S. Staab. Learning concept hierarchies from text with a guided agglomerative clustering algorithm. In *Proceedings of the ICML 2005 Workshop on Learning and Extending Lexical Ontologies with Machine Learning Methods*, 2005.

[40] P. Cimiano and J. Völker. text2onto. In *International conference on application of natural language to information systems*, pages 227–238. Springer, 2005.

[41] J. Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

[42] G. Contissa, K. Docter, F. Lagioia, M. Lippi, H.-W. Micklitz, P. Pałka, G. Sartor, and P. Torroni. Claudette meets gdpr: Automating the evaluation of privacy policies using artificial intelligence. *Available at SSRN 3208596*, 2018.

[43] Council od Europe. The modernised convention 108: novelties in a nutshell. https://rm.coe.int/modernised-conv-overview-of-the-novelties/16808accf8. Accessed: 2020-08-18.

[44] Council of Europe and European Union Agency for Fundamental Rights. Context and background of european data protection law. In *Handbook on European data protection law*, chapter 1, pages 15–80. Publications Office of the European Union, 2018.

[45] Council of Europe and European Union Agency for Fundamental Rights. Key principles of european data protection law. In *Handbook on European data protection law*, chapter 3, pages 115–138. Publications Office of the European Union, 2018.

[46] E. Daga, V. Presutti, A. Gangemi, and A. Salvati. http://ontologydesignpatterns.org [ODP]. In *Proceedings of the 2007 International Conference on Posters and Demonstrations*, volume 401 of *ISWC-PD'08*, page 169–170. CEUR Workshops Proceedings, 2008.

[47] D. Das, N. Schneider, D. Chen, and N. A. Smith. Semafor 1.0: A probabilistic frame-semantic parser. *Language Technologies Institute, School of Computer Science, Carnegie Mellon University*, 2010.

[48] R. Davis, H. Shrobe, and P. Szolovits. What is a knowledge representation? *AI magazine*, 14(1):17–17, 1993.

[49] A. De Nicola and M. Missikoff. A lightweight methodology for rapid ontology engineering. *Communications of the ACM*, 59(3):79–86, 2016.

[50] C. M. de Oliveira Rodrigues, F. L. G. de Freitas, E. F. S. Barreiros, R. R. de Azevedo, and A. T. de Almeida Filho. Legal ontologies over time: A

systematic mapping study. *Expert Systems with Applications*, 130:12–30, 2019.

[51] C. de Terwangne. *The Right to be Forgotten and Informational Autonomy in the Digital Environment*, pages 82–101. Palgrave Macmillan UK, London, 2014.

[52] M. De Vos, S. Kirrane, J. Padget, and K. Satoh. Odrl policy modelling and compliance checking. In P. Fodor, M. Montali, D. Calvanese, and D. Roman, editors, *Rules and Reasoning*, pages 36–51. Springer International Publishing, 2019.

[53] C. Debruyne, H. J. Pandit, D. Lewis, and D. O'Sullivan. "just-in-time" generation of datasets by considering structured representations of given consent for gdpr compliance. *KNOWLEDGE AND INFORMATION SYSTEMS*, 2020.

[54] A. Degbelo. A snapshot of ontology evaluation criteria and strategies. In *Proceedings of the 13th International Conference on Semantic Systems*, Semantics2017, page 1–8, New York, NY, USA, 2017. Association for Computing Machinery.

[55] L. Del Corro and R. Gemulla. Clausie: Clause-based open information extraction. In *Proceedings of the 22nd International Conference on World Wide Web*, WWW '13, page 355–366, New York, NY, USA, 2013. Association for Computing Machinery.

[56] C. Delli Bovi, L. Telesca, and R. Navigli. Large-scale information extraction from textual definitions through deep syntactic and semantic analysis. *Transactions of the Association for Computational Linguistics*, 3:529–543, 2015.

[57] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[58] L. Ding, P. Kolari, Z. Ding, and S. Avancha. Using ontologies in the semantic web: A survey. In *Ontologies*, pages 79–113. Springer, 2007.

[59] I. Distinto, M. d'Aquin, and E. Motta. Loted2: An ontology of european public procurement notices. *Semantic Web*, 7(3):267–293, 2016.

[60] K. Doing-Harris, Y. Livnat, and S. Meystre. Automated concept and relationship extraction for the semi-automated ontology management (seam) system. *Journal of biomedical semantics*, 6(1):15, 2015.

[61] E. Drymonas, K. Zervanou, and E. G. Petrakis. Unsupervised ontology acquisition from plain texts: the ontogain system. In *International Conference on Application of Natural Language to Information Systems*, pages 277–287. Springer, 2010.

[62] J. Elhassouni, A. El qadi, Y. El madani El alami, and M. El haziti. Modeling with ontologies design patterns: Credit risk scorecard as a case study. *Indonesian Journal of Electrical Engineering and Computer Science*, 17(1):429–439, 2020.

[63] T. Endicott. Law is necessarily vague. *LEG*, 7:379, 2001.

[64] T. Ermakova, A. Baumann, B. Fabian, and H. Krasnova. Privacy policies and users' trust: Does readability matter? In *Americas Conference on Information Systems*, 2014.

[65] B. D. Eugenio and M. Glass. The kappa statistic: A second look. *Computational linguistics*, 30(1):95–101, 2004.

[66] European Commission - Official Website. What are data protection authorities (dpas)? `https://ec.europa.eu/info/law/law-topic/data-protection/reform/what-are-data-protection-authorities-dpas_en`. Accessed: 2020-08-21.

[67] European Commission - Official Website. Why do we need the charter? `https://ec.europa.eu/info/aid-development-cooperation-fundamental-rights/your-rights-eu/eu-charter-fundamental-rights/why-do-we-need-charter_it`. Accessed: 2020-08-18.

[68] European Commission - Official Website. European commission sets out strategy to strengthen eu data protection rules. `https://ec.europa.eu/commission/presscorner/detail/en/IP_10_1462`, 2010. Accessed: 2020-08-18.

[69] European Commission - Press release. Data protection regulation one year on: 73% of europeans have heard of at least one of their rights. `https://ec.europa.eu/commission/presscorner/detail/en/IP_19_2956`, 2019. Accessed: 2020-08-16.

[70] B. Fabian, T. Ermakova, and T. Lentz. Large-scale readability analysis of privacy policies. In *Proceedings of the International Conference on Web Intelligence*, pages 18–25, 2017.

[71] A. Fader, S. Soderland, and O. Etzioni. Identifying relations for open information extraction. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1535–1545, Edinburgh, Scotland, UK., jul 2011. Association for Computational Linguistics.

[72] R. Fanguy, B. Kleen, and L. Soule. Privacy policies: cloze test reveals readability concerns. *Issues in Information Systems*, 5(1):117–123, 2004.

[73] M. Fernandez-Lopez, A. Gomez-Perez, and N. Juristo. Methontology: from ontological art towards ontological engineering. In *Proceedings of the AAAI97 Spring Symposium*, pages 33–40, 1997.

[74] C. J. Fillmore et al. Frame semantics and the nature of language. In *Annals of the New York Academy of Sciences: Conference on the origin and development of language and speech*, volume 280, pages 20–32. New York, 1976.

[75] R. Flesch. A new readability yardstick. *Journal of applied psychology*, 32(3):221, 1948.

[76] T. Francart, J. DANN, R. Pappalardo, C. Malagon, and M. Pellegrino. The european legislation identifier. *Knowledge of the Law in the Big Data Age*, 317:137, 2019.

[77] K. Frantzi, S. Ananiadou, and H. Mima. Automatic recognition of multi-word terms:. the c-value/nc-value method. *International journal on digital libraries*, 3(2):115–130, 2000.

[78] M. Gallé, A. Christofi, and H. Elsahar. The case for a GDPR-specific annotated dataset of privacy policies. In W. Shomir, S. Ghanavati, K. Ghazinour, and N. Sadeh, editors, *Proceedings of the PAL: Privacy-Enhancing Artificial Intelligence and Language Technologies As Part of the AAAI Spring Symposium Series (AAAI-SSS 2019)*, volume 2335. CEUR Workshop Proceedings, 2019.

[79] F. Gandon, G. Governatori, and S. Villata. Normative requirements as linked data. In A. Wyner and G. Casini, editors, *Legal Knowledge and Information Systems*, volume 302 of *Frontiers in Artificial Intelligence and Applications*, pages 1–10. IOS Press, 2017.

[80] A. Gangemi. Ontology design patterns for semantic web content. In *International semantic web conference*, pages 262–276. Springer, 2005.

[81] A. Gangemi, C. Catenacci, M. Ciaramita, and J. Lehmann. Modelling ontology evaluation and validation. In *European Semantic Web Conference*, pages 140–154. Springer, 2006.

[82] A. Gangemi and V. Presutti. Ontology design patterns. In *Handbook on ontologies*, pages 221–243. Springer, 2009.

[83] A. Gangemi, M.-T. Sagri, and D. Tiscornia. A constructive framework for legal ontologies. In *Law and the semantic web*, pages 97–124. Springer, 2005.

[84] S. González-Carvajal and E. C. Garrido-Merchán. Comparing bert against traditional machine learning text classification. *arXiv preprint arXiv:2005.13012*, 2020.

[85] T. R. Gruber. Ontolingua: A mechanism to support portable ontologies, 1992.

[86] T. R. Gruber. Toward principles for the design of ontologies used for knowledge sharing? *International Journal of Human-Computer Studies*, 43(5):907 – 928, 1995.

[87] N. Guarino. *Formal Ontology in Information Systems: Proceedings of the 1st International Conference June 6-8, 1998, Trento, Italy*. IOS Press, NLD, 1st edition, 1998.

[88] N. Guarino and P. Giaretta. Ontologies and knowledge bases: Towards a terminological clarification. In *Towards Very Large Knowledge Bases: Knowledge Building & Knowledge Sharing*, pages 25–32. IOS Press, 1995.

[89] N. Guarino and C. Welty. Evaluating ontological decisions with ontoclean. *Communications of the ACM*, 45(2):61–65, 2002.

[90] M. Guerrini and T. Possemato. Linked data: a new alphabet for the semantic web. *JLIS. it*, 4(1):67, 2013.

[91] N. Gupta, S. Podder, S. Sengupta, and K. Annervaz. Domain ontology induction using word embeddings. In *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 115–119. IEEE, 2016.

[92] K. Hammar. *Content Ontology Design Patterns: Qualities, Methods, and Tools. PhD Thesis*. 2017.

[93] H. Harkous, K. Fawaz, R. Lebret, F. Schaub, K. G. Shin, and K. Aberer. Polisis: Automated analysis and presentation of privacy policies using deep learning. In *27th {USENIX} Security Symposium ({USENIX} Security 18)*, pages 531–548, 2018.

[94] M. Hazman, S. R. El-Beltagy, and A. Rafea. A survey of ontology learning approaches. *International Journal of Computer Applications*, 22(8):36–43, May 2011.

[95] M. A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *COLING 1992 Volume 2: The 15th International Conference on Computational Linguistics*, 1992.

[96] M. Hepp. Possible ontologies: How reality constrains the development of relevant ontologies. *IEEE Internet Computing*, 11(1):90–96, 2007.

[97] M. Hepp. Ontologies: State of the art, business potential, and grand challenges. In *Ontology Management*, pages 3–22. Springer, 2008.

[98] A. R. Hippisley, D. Cheng, and K. Ahmad. The head-modifier principle and multilingual term extraction. *Natural Language Engineering*, 11(2):129, 2005.

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

[100] R. Hoekstra, J. Breuker, M. Di Bello, A. Boer, et al. The lkif core ontology of basic legal concepts. *LOAIT*, 321:43–63, 2007.

[101] I. Horrocks et al. Daml+oil: A description logic for the semantic web. *IEEE Data Eng. Bull.*, 25(1):4–9, 2002.

[102] L. Humphreys, G. Boella, L. Di Caro, L. Robaldo, L. van der Torre, S. Ghanavati, and R. Muthuri. Populating legal ontologies using semantic role labeling. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2157–2166, Marseille, France, May 2020. European Language Resources Association.

[103] C. Jensen and C. Potts. Privacy policies as decision-making tools: an evaluation of online privacy notices. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 471–478, 2004.

[104] D. Jurafsky and J. H. Martin. Information Extraction. Draft of October 2019. In *Speech and Language Processing*. 2019.

[105] D. Jurafsky and J. H. Martin. Semantic Role Labelling. Draft of October 2019. In *Speech and Language Processing*. 2019.

[106] D. D. Kehagias, I. Papadimitriou, J. Hois, D. Tzovaras, and J. Bateman. A methodological approach for ontology evaluation and refinement. In *ASK-IT Final Conference. June.(Cit. on p.)*, pages 1–13, 2008.

[107] M. Keymanesh, M. Elsner, and S. Parthasarathy. Toward domain-guided controllable summarization of privacy policies.

[108] K. Krasnashchok, M. Mustapha, A. Al Bassit, and S. Skhiri. Towards privacy policy conceptual modeling. In *International Conference on Conceptual Modeling*, pages 429–438. Springer, 2020.

[109] B. Krumay and J. Klar. Readability of privacy policies. In *IFIP Annual Conference on Data and Applications Security and Privacy*, pages 388–399. Springer, 2020.

[110] G. Lame. Using nlp techniques to identify legal ontology components: concepts and relations. In *Law and the Semantic Web*, pages 169–184. Springer, 2005.

[111] J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174, 1977.

[112] J. Lehmann, S. Auer, L. Bühmann, and S. Tramp. Class expression learning for ontology engineering. *Journal of Web Semantics*, 9(1):71 – 81, 2011.

[113] V. Leone, G. Siragusa, L. Di Caro, and R. Navigli. Building semantic grams of human knowledge. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2991–3000, Marseille, France, May 2020. European Language Resources Association.

[114] B. Levin. *English verb classes and alternations: A preliminary investigation*. University of Chicago press, 1993.

[115] D. Liebwald. Law's capacity for vagueness. *International Journal for the Semiotics of Law-Revue internationale de Sémiotique juridique*, 26(2):391–423, 2013.

[116] T. Linden, R. Khandelwal, H. Harkous, and K. Fawaz. The privacy policy landscape after the gdpr. *Proceedings on Privacy Enhancing Technologies*, 2020(1):47–64, 2020.

[117] K. Litman-Navarro. We read 150 privacy policies. they were an incomprehensible disaster. https://www.nytimes.com/interactive/2019/06/12/opinion/facebook-google-privacy-policies.html, 2019. Accessed: 2020-08-11.

[118] S. Lodhi and Z. Ahmed. *Content Ontology Design Pattern Presentation. Master Thesis*. 2011.

[119] M. Lubani, S. A. M. Noah, and R. Mahmud. Ontology population: approaches and design aspects. *Journal of Information Science*, 45(4):502–515, 2019.

[120] A. Maedche and S. Staab. Ontology learning for the semantic web. *IEEE Intelligent Systems*, 16(2):72–79, 2001.

[121] J. Malone and H. Parkinson. Reference and application ontologies. *Ontogenesis*, 2010.

[122] L. Màrquez, X. Carreras, K. C. Litkowski, and S. Stevenson. Special issue introduction: Semantic role labeling: An introduction to the special issue. *Computational Linguistics*, 34(2):145–159, 2008.

[123] A. K. Massey, R. L. Rutledge, A. I. Antón, and P. P. Swire. Identifying and classifying ambiguity for regulatory requirements. In *2014 IEEE 22nd international requirements engineering conference (RE)*, pages 83–92. IEEE, 2014.

[124] Mausam, M. Schmitz, S. Soderland, R. Bart, and O. Etzioni. Open language learning for information extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 523–534, Jeju Island, Korea, July 2012. Association for Computational Linguistics.

[125] A. M. McDonald and L. F. Cranor. The cost of reading privacy policies. *Isjlp*, 4:543, 2008.

[126] D. L. McGuinness. Ontologies come of age. *Spinning the semantic web: bringing the World Wide Web to its full potential*, pages 171–194, 2002.

[127] M. L. McHugh. Interrater reliability: the kappa statistic. *Biochemia medica: Biochemia medica*, 22(3):276–282, 2012.

[128] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[129] G. A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, 1995.

[130] G. R. Milne and M. J. Culnan. Strategies for reducing online privacy risks: Why consumers read (or don't read) online privacy notices. *Journal of interactive marketing*, 18(3):15–29, 2004.

[131] F. Moerdijk et al. Frames and semagrams. meaning description in the general dutch dictionary. In *Proceedings of the Thirteenth Euralex International Congress, EURALEX 2008*, 2008.

[132] E. Montiel-Ponsoda, J. Gracia, and V. Rodríguez-Doncel. Building the legal knowledge graph for smart compliance services in multilingual europe. In *Proceedings of the 1st Workshop on Technologies for Regulatory Compliance co-located with the 30th International Conference on Legal Knowledge and Information Systems (JURIX 2017)*, pages 15–17, 2017.

[133] A. Moro and R. Navigli. Integrating syntactic and semantic analysis into the open information extraction paradigm. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, IJCAI '13, page 2148–2154. AAAI Press, 2013.

[134] A. Moses and G. B. Kumar. Snippet generation using textbook corpus-an nlp approach based on bert. In *Journal of Physics: Conference Series*, volume 1716, page 012061. IOP Publishing, 2020.

[135] J. F. Muñoz-Soro, G. Esteban, O. Corcho, and F. Serón. Pproc, an ontology for transparency in public procurement. *Semantic Web*, 7(3):295–309, 2016.

[136] R. Nanda, G. Siragusa, L. Di Caro, M. Theobald, G. Boella, L. Robaldo, and F. Costamagna. Concept recognition in european and national law. In *JURIX*, pages 193–198, 2017.

[137] R. Navigli. Ontologies. In R. Mitkov, editor, *The Oxford Handbook of Computational Linguistics 2nd edition*, Oxford, United Kingdom, 2016. Oxford University Press.

[138] R. Navigli and S. P. Ponzetto. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217 – 250, 2012.

[139] R. Navigli, P. Velardi, and A. Gangemi. Ontology learning and its application to automated terminology translation. *IEEE Intelligent systems*, 18(1):22–31, 2003.

[140] N. M. Nejad, D. Graux, and D. Collarana. Towards measuring risk factors in privacy policies. In *Proceedings of the Workshop on Artificial Intelligence and the Administrative State co-located with 17th International Conference on AI and Law (ICAIL 2019)*, volume 2471, pages 18–20. CEUR Workshop Proceedings, 2019.

[141] N. M. Nejad, P. Jabat, R. Nedelchev, S. Scerri, and D. Graux. Establishing a strong baseline for privacy policy classification. In *IFIP International Conference on ICT Systems Security and Privacy Protection*, 2020.

[142] N. M. Nejad, S. Scerri, and J. Lehmann. Knight: Mapping privacy policies to gdpr. In *European Knowledge Acquisition Workshop*, pages 258–272. Springer, 2018.

[143] C. Niklaus, M. Cetto, A. Freitas, and S. Handschuh. A survey on open information extraction. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3866–3878, Santa Fe, New Mexico, USA, Aug. 2018. Association for Computational Linguistics.

[144] H. Nissenbaum. *Privacy in context: Technology, policy, and the integrity of social life*. Stanford University Press, 2009.

[145] N. F. Noy, D. L. McGuinness, et al. Ontology development 101: A guide to creating your first ontology, 2001.

[146] L. Obrst, W. Ceusters, I. Mani, S. Ray, and B. Smith. The evaluation of ontologies. In *Semantic web*, pages 139–158. Springer, 2007.

[147] M. Palmirani, G. Bincoletto, V. Leone, S. Sapienza, and F. Sovrano. Hybrid refining approach of pronto ontology. In A. Kő, E. Francesconi, G. Kotsis, A. M. Tjoa, and I. Khalil, editors, *Electronic Government and the Information Systems Perspective*, pages 3–17, Cham, 2020. Springer International Publishing.

[148] M. Palmirani, G. Governatori, A. Rotolo, S. Tabet, H. Boley, and A. Paschke. Legalruleml: Xml-based rules and norms. In *Rule-Based Modeling and Computing on the Semantic Web*, pages 298–312. Springer, 2011.

[149] M. Palmirani, M. Martoni, A. Rossi, C. Bartolini, and L. Robaldo. Legal ontology for modelling gdpr concepts and norms. In *JURIX*, pages 91–100, 2018.

[150] M. Palmirani, M. Martoni, A. Rossi, C. Bartolini, and L. Robaldo. Pronto: Privacy ontology for legal compliance. In *ECDG 2018 18th European Conference on Digital Government*, page 142. Academic Conferences and publishing limited, 2018.

[151] A. Panchenko, S. Faralli, E. Ruppert, S. Remus, H. Naets, C. Fairon, S. P. Ponzetto, and C. Biemann. Taxi at semeval-2016 task 13: a taxonomy induction method based on lexico-syntactic patterns, substrings and focused crawling. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1320–1327, 2016.

[152] H. J. Pandit, C. Debruyne, D. O'Sullivan, and D. Lewis. Gconsent - a consent ontology based on the gdpr. In P. Hitzler, M. Fernández, K. Janowicz, A. Zaveri, A. J. Gray, V. Lopez, A. Haller, and K. Hammar, editors, *The Semantic Web*, pages 270–282, Cham, 2019. Springer International Publishing.

[153] H. J. Pandit, K. Fatema, D. O'Sullivan, and D. Lewis. Gdprtext-gdpr as a linked data resource. In *European Semantic Web Conference*, pages 481–495. Springer, 2018.

[154] H. J. Pandit and D. Lewis. Modelling provenance for gdpr compliance using linked open data vocabularies. In C. Brewster, M. Cheatham, M. d'Aquin, S. Decker, and S. Kirrane, editors, *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)*, number 1951 in CEUR Workshop Proceedings, pages 481–495. Springer, 2017.

[155] H. J. Pandit, D. O'Sullivan, and D. Lewis. An ontology design pattern for describing personal data in privacy policies. In M. G. Skjæveland, Y. Hu, K. Hammar, V. Svátek, and A. Ławrynowicz, editors, *Proceedings of the 9th Workshop on Ontology Design and Patterns (WOP 2018) co-located with 17th International Semantic Web Conference (ISWC 2018)*, number 2195 in CEUR Workshop Proceedings, pages 29–39. Springer, 2018.

[156] H. J. Pandit, A. Polleres, B. Bos, R. Brennan, B. Bruegger, F. J. Ekaputra, J. D. Fernández, R. G. Hamed, E. Kiesling, M. Lizar, E. Schlehahn, S. Steyskal, and R. Wenning. Creating a vocabulary for data privacy. In H. Panetto, C. Debruyne, M. Hepp, D. Lewis, C. A. Ardagna, and R. Meersman, editors, *On the Move to Meaningful Internet Systems: OTM 2019 Conferences*, pages 714–730, Cham, 2019. Springer International Publishing.

[157] İ. Pembeci. Using word embeddings for ontology enrichment. *International Journal of Intelligent Systems and Applications in Engineering*, 4(3):49–56, 2016.

[158] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[159] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, 2018. Association for Computational Linguistics.

[160] H. S. Pinto, S. Staab, C. Tempich, and Y. Sure. *Distributed Engineering of Ontologies (DILIGENT)*, pages 303–322. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006.

[161] I. Pollach. A typology of communicative strategies in online privacy policies: Ethics, power and informed consent. *Journal of Business Ethics*, 62(3):221, 2005.

[162] I. Pollach. What's wrong with online privacy policies? *Communications of the ACM*, 50(9):103–108, 2007.

[163] V. Presutti, A. Gangemi, S. David, G. A. de Cea, M. Suárez-Figueroa, E. Montiel-Ponsoda, and M. Poveda. A library of ontology design patterns: reusable solutions for collaborative design of networked ontologies. NeOn Project Deliverable D2.5.1, 2008.

[164] C. Qu, L. Yang, M. Qiu, W. B. Croft, Y. Zhang, and M. Iyyer. Bert with history answer embedding for conversational question answering. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1133–1136, 2019.

[165] J. Raad and C. Cruz. A survey on ontology evaluation methods. In *Proceedings of the International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*, IC3K 2015, page 179–186. SCITEPRESS - Science and Technology Publications, Lda, 2015.

[166] J. R. Reidenberg, J. Bhatia, T. D. Breaux, and T. B. Norton. Ambiguity in privacy policies and the impact of regulation. *The Journal of Legal Studies*, 45(S2):S163–S190, 2016.

[167] L. Robaldo, C. Bartolini, M. Palmirani, A. Rossi, M. Martoni, and G. Lenzini. Formalizing gdpr provisions in reified i/o logic: the dapreco knowledge base. *Journal of Logic, Language and Information*, pages 1–49, 2019.

[168] C. Roussey, F. Pinet, M. A. Kang, and O. Corcho. An introduction to ontologies and ontology engineering. In *Ontologies in Urban development projects*, pages 9–38. Springer, 2011.

[169] P. Ryan, M. Crane, and R. Brennan. Design challenges for gdpr regtech. *arXiv preprint arXiv:2005.12138*, 2020.

[170] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing Management*, 24(5):513 – 523, 1988.

[171] D. Sarne, J. Schler, A. Singer, A. Sela, and I. Bar Siman Tov. Unsupervised topic extraction from privacy policies. In *Companion Proceedings of The 2019 World Wide Web Conference*, pages 563–568, 2019.

[172] A. Sattar, E. S. M. Surin, M. N. Ahmad, M. Ahmad, and A. K. Mahmood. Comparative analysis of methodologies for domain ontology development: A systematic review. *International Journal of Advanced Computer Science and Applications*, 11(5), 2020.

[173] S. Schaffert, A. Gruber, and R. Westenthaler. A semantic wiki for collaborative knowledge formation. In *SEMANTICS 2005*, Vienna, Austria, 2005.

[174] P.-N. Schwab. Reading privacy policies of the 20 most-used mobile apps takes 6h40. `https://www.intotheminds.com/blog/en/reading-privacy-policies-of-the-20-most-used-mobile-apps-takes-6h40/`, 2018. Accessed: 2020-08-16.

[175] M. Shamsfard and A. A. Barforoush. The state of the art in ontology learning: a framework for comparison. *The Knowledge Engineering Review*, 18(4):293, 2003.

[176] Y. Shvartzshanider, A. Balashankar, T. Wies, and L. Subramanian. Recipe: Applying open domain question answering to privacy policies. In *Proceedings of the Workshop on Machine Reading for Question Answering*, pages 71–77, 2018.

[177] Y. Shvartzshnaider, N. Apthorpe, N. Feamster, and H. Nissenbaum. Going against the (appropriate) flow: a contextual integrity approach to

privacy policy analysis. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7, pages 162–170, 2019.

[178] E. P. B. Simperl and C. Tempich. Ontology engineering: A reality check. In R. Meersman and Z. Tari, editors, *On the Move to Meaningful Internet Systems 2006: CoopIS, DOA, GADA, and ODBASE*, pages 836–854, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.

[179] R. I. Singh, M. Sumeeth, and J. Miller. A user-centric evaluation of the readability of privacy policies in popular web sites. *Information Systems Frontiers*, 13(4):501–514, 2011.

[180] G. Siragusa, R. Nanda, V. De Paiva, and L. Di Caro. Relating legal entities via open information extraction. In *Research Conference on Metadata and Semantics Research*, pages 181–187. Springer, 2018.

[181] B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, L. J. Goldberg, K. Eilbeck, A. Ireland, C. J. Mungall, et al. The obo foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature biotechnology*, 25(11):1251–1255, 2007.

[182] B. Smith and C. Welty. Ontology: Towards a new synthesis. In *Formal Ontology in Information Systems*, volume 10, pages 3–9. ACM Press, 2001.

[183] M. Sordo, S. Oramas, and L. Espinosa-Anke. Extracting relations from unstructured text sources for music recommendation. In *International Conference on Applications of Natural Language to Information Systems*, pages 369–382. Springer, 2015.

[184] K. T. Stevens, K. C. Stevens, and W. P. Stevens. Measuring the readability of business writing: The cloze procedure versus readability formulas. *The Journal of Business Communication (1973)*, 29(4):367–382, 1992.

[185] M. C. Suárez-Figueroa, A. Gómez-Pérez, and M. Fernández-López. *The NeOn Methodology for Ontology Engineering*, pages 9–34. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[186] M. Sumeeth, R. Singh, and J. Miller. Are online privacy policies readable? *International Journal of Information Security and Privacy (IJISP)*, 4(1):93–116, 2010.

[187] W. B. Tesfay, P. Hofmann, T. Nakamura, S. Kiyomoto, and J. Serna. Privacyguide: towards an implementation of the eu gdpr on internet

privacy policy evaluation. In *Proceedings of the Fourth ACM International Workshop on Security and Privacy Analytics*, pages 15–21, 2018.

[188] D. Turcato, F. Popowich, J. Toole, D. Fass, D. Nicholson, and G. Tisher. Adapting a synonym database to specific domains. In *ACL-2000 Workshop on Recent Advances in Natural Language Processing and Information Retrieval*, pages 1–11, 2000.

[189] A. Valente, J. Breuker, et al. A functional ontology of law. *Towards a global expert system in law*, pages 112–136, 1994.

[190] R. Van Gog and T. M. Van Engers. Modeling legislation using natural language processing. In *2001 IEEE International Conference on Systems, Man and Cybernetics. e-Systems and e-Man for Cybernetics in Cyberspace (Cat. No. 01CH37236)*, volume 1, pages 561–566. IEEE, 2001.

[191] P. Voigt and A. Von dem Bussche. Rights of data subjects. In *The EU general data protection regulation (GDPR)*, chapter 5, pages 141–188. Springer International Publishing, 1 edition, 2017.

[192] P. Voigt and A. Von dem Bussche. Scope of application of the gdpr. In *The EU general data protection regulation (GDPR)*, chapter 2, pages 9–30. Springer International Publishing, 1 edition, 2017.

[193] J. Völker, D. Vrandečić, Y. Sure, and A. Hotho. Learning disjointness. In E. Franconi, M. Kifer, and W. May, editors, *The Semantic Web: Research and Applications*, pages 175–189, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg.

[194] D. Vrandečić. Ontology evaluation. In *Handbook on ontologies*, pages 293–313. Springer, 2009.

[195] M. J. Warrens. Inequalities between multi-rater kappas. *Advances in data analysis and classification*, 4(4):271–286, 2010.

[196] S. Wilson, F. Schaub, A. A. Dara, F. Liu, S. Cherivirala, P. Giovanni Leon, M. Schaarup Andersen, S. Zimmeck, K. M. Sathyendra, N. C. Russell, T. B. Norton, E. Hovy, J. Reidenberg, and N. Sadeh. The creation and analysis of a website privacy policy corpus. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1330–1340, Berlin, Germany, Aug. 2016. Association for Computational Linguistics.

[197] W. Wong, W. Liu, and M. Bennamoun. Ontology learning from text: A look back and into the future. *ACM Computing Surveys (CSUR)*, 44(4):1–36, 2012.

[198] M. Yahya, S. Whang, R. Gupta, and A. Y. Halevy. Renoun: Fact extraction for nominal attributes. In A. Moschitti, B. Pang, and W. Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 325–335. ACL, 2014.

[199] R. N. Zaeem, R. L. German, and K. S. Barber. Privacycheck: Automatic summarization of privacy policies using data mining. *ACM Transactions on Internet Technology*, 18(4), 2018.

[200] G. Zanfir-Fortuna. Chapter III rights of the data subject (articles 12-23). article 13. information to be provided where personal data are collected from the data subject. In *General Data Protection Regulation. Article-by-Article Commentary*, pages 413–433. Hart Publishing, 2020.

[201] M. L. Zeng. Knowledge organization systems (kos). *KO KNOWLEDGE ORGANIZATION*, 35(2-3):160–182, 2008.