

**Digital Ecosystems and Business Intelligence Institute**

**A Generic Privacy Ontology and  
its Applications to Different Domains**

**Michael Hecker**

**This thesis is presented for the Degree of  
Doctor of Philosophy  
of  
Curtin University of Technology**

**March 2009**

Declaration

To the best of my knowledge and belief this thesis contains no material previously published by any other person except where due acknowledgment has been made.

This thesis contains no material which has been accepted for the award of any other degree or diploma in any university.

Signature: .....

Date: .....

Table of Figures.....	9
Summary of thesis.....	11
1. Introduction.....	13
1.1. Broad overview of privacy.....	13
1.2. Examples .....	14
1.3. Motivation.....	18
1.4. Privacy by Design .....	18
1.5. Scope of problem.....	19
1.6. Plan of thesis .....	21
2. Review and Evaluation of literature.....	23
2.1. Privacy - A Historical View and its Evolution.....	23
2.2. Different Notions of Privacy .....	24
2.3. Privacy and Legislation .....	26
2.4. Privacy issues and challenges on the web .....	28
2.5. Privacy preserving techniques .....	29
2.6. Privacy issues and challenges on the semantic web.....	31
2.6.1. Semantic and ontological concepts.....	32
2.7. Privacy preserving techniques in Data Mining.....	35
2.8. Summary .....	36
3. Problem definition .....	37
3.1. Terms and concepts used .....	37
3.1.1. Privacy.....	37
3.1.1.1. The right data .....	38
3.1.1.2. The right purpose.....	38
3.1.1.3. What about Confidentiality? .....	39
3.1.1.4. Definition of Privacy.....	39
3.1.2. Trust & Reputation in Privacy .....	40
3.1.3. Security & Safeguards in Privacy.....	41
3.1.4. Entity.....	42
3.1.5. Data Subject .....	43
3.1.6. Resource.....	43
3.1.7. Personal Information .....	44
3.1.8. Identity.....	44

3.1.9.	Process .....	45
3.1.10.	Policy.....	45
3.1.11.	Consent .....	45
3.1.12.	Repository .....	46
3.2.	Privacy Principles .....	46
3.2.1.	Data Quality.....	46
3.2.2.	Transparency .....	47
3.2.3.	Intention and Notification .....	47
3.2.4.	Finality Principle .....	47
3.2.5.	Legitimate Grounds of Processing.....	48
3.2.6.	Data Subject's rights.....	48
3.2.7.	Security .....	49
3.2.8.	Accountability.....	49
3.2.9.	Openness.....	49
3.2.10.	Anonymity .....	50
3.2.11.	Transfer of personal information between different jurisdictions .....	50
3.3.	Problem definition .....	50
3.3.1.	Summary of problem definition.....	51
3.4.	Choice of methodology to problem solving .....	52
3.4.1.	Natural language processing.....	52
3.4.2.	Mathematical representation .....	53
3.4.3.	Knowledge representation .....	54
3.4.4.	Glossary of terms .....	54
3.4.4.1.	Topic Maps .....	55
3.4.4.2.	Ontologies .....	55
4.	Solution Overview / Roadmap.....	57
4.1.	Ontology development.....	57
4.2.	Ontology architecture .....	58
4.3.	Conclusion .....	67
5.	Generic Privacy Ontology - in details.....	68
5.1.	Introduction.....	68
5.2.	Ontology concepts .....	68
5.2.1.	Entity hierarchy.....	70

5.2.2.	Resources .....	75
5.3.	Privacy Processes .....	78
5.4.	Detailed illustration .....	80
5.5.	The Privacy Principles.....	81
5.5.1.	Data Quality (Quality Aspect 1).....	82
5.5.1.1.	Adequate (Quality assessment criteria 1).....	82
5.5.1.2.	Relevance to purpose (Quality assessment criteria 2).....	82
5.5.1.3.	Correctness (Quality assessment criteria 3).....	82
5.5.2.	Security (Quality Aspect 2) .....	83
5.5.2.1.	Safeguards adequate (Quality assessment criteria 1).....	83
5.5.2.2.	Security policy adequate .....	83
5.5.2.3.	Data destruction policy adequate .....	83
5.5.2.4.	Contingency plan adequate .....	83
5.5.2.5.	Personnel requirements adequate.....	83
5.5.2.6.	Privacy enhancing technologies adequate.....	83
5.5.2.7.	ICT infrastructure adequate.....	83
5.5.3.	Data subject's rights .....	84
5.5.3.1.	Access privileges to own data .....	84
5.5.3.2.	Level of ability to request rectifications/supplementations/deletions.....	84
5.5.3.3.	Ability to block content for certain purposes.....	84
5.5.3.4.	Ability to object against processing.....	84
5.5.4.	Legitimate Grounds of Processing.....	84
5.5.4.1.	Unambiguous consent.....	84
5.5.4.2.	Processing to fulfill contract requirement .....	84
5.5.4.3.	Legal reasons .....	85
5.5.4.4.	Protection of vital interest of data subject .....	85
5.5.4.5.	Data belongs to following sensitive category .....	85
5.5.5.	Transparency .....	85
5.5.5.1.	Data obtained directly notify of processing.....	85
5.5.5.2.	Data obtained indirectly notify of processing .....	85
5.5.5.3.	Identity of processor revealed .....	85
5.5.5.4.	Purpose of data stated .....	86
5.5.5.5.	Recording in accordance with law.....	86

5.5.5.6.	Third party involved .....	86
5.5.5.7.	Ability to object to involve third party .....	86
5.5.6.	Finality principle .....	86
5.5.6.1.	Level of purpose specified .....	86
5.5.6.2.	Purpose legitimate .....	86
5.5.6.3.	Retention period.....	86
5.5.7.	Processing by a third party - data sharing.....	87
5.5.7.1.	Instructed by controller.....	87
5.5.7.2.	Level of compliance with obligations of controller .....	87
5.5.7.3.	Legal binding contract in place.....	87
5.5.8.	Accountability.....	87
5.5.8.1.	Person nominated to watch over compliance.....	87
5.5.9.	Openness.....	87
5.5.9.1.	Policies about procedures available.....	87
5.5.10.	Anonymity .....	87
5.5.10.1.	Data anonymized .....	87
5.5.11.	Consent .....	87
5.5.11.1.	Explicit.....	88
5.5.11.2.	Implicit .....	88
5.5.12.	Transfer between different jurisdictions .....	88
5.5.12.1.	If transfer - justification to transfer data to jurisdiction with different privacy protection laws .....	88
5.5.12.2.	Privacy Protection Laws <i>Standards</i> .....	88
5.6.	Relationship between concepts and principles.....	89
5.7.	Privacy evaluation process .....	90
5.8.	Conclusion .....	93
6.	Specialization I - Restricted medical domain.....	94
6.1.	Introduction.....	94
6.2.	Medical Domain Ontology.....	94
6.2.1.	Concepts.....	95
6.2.2.	Processes .....	97
6.3.	Medical Privacy Ontology.....	99
6.4.	Example of instances.....	106

6.5.	Evaluation.....	109
6.6.	Conclusion .....	111
7.	Specialization II - B2C E-Commerce domain.....	113
7.1.	Introduction.....	113
7.2.	Restricted B2C E-Commerce Domain Ontology.....	114
7.2.1.	Basic high level concepts .....	114
7.2.2.	Technical level concepts .....	118
7.2.3.	Technical level processes.....	121
7.3.	E-Commerce Privacy Ontology.....	123
7.4.	Example of instances.....	129
7.4.1.	Concepts.....	129
7.4.2.	Processes.....	133
7.5.	Evaluation of the abstract technical level.....	135
7.6.	Conclusion .....	137
8.	Implementation .....	138
8.1.	Introduction.....	138
8.2.	Choice of languages.....	138
8.2.1.	RDF and RDFS .....	138
8.2.2.	OWL .....	139
8.3.	Tools.....	141
8.3.1.	Protégé.....	141
8.4.	Examples (screenshots and code snippets).....	141
8.4.1.	Main concepts .....	141
8.4.2.	Privacy process .....	146
8.4.3.	Privacy principles .....	147
8.4.4.	Quality assessment criteria values.....	150
8.5.	Conclusion .....	153
9.	Summary of thesis.....	155
9.1.	Introduction.....	155
9.2.	Recapitulation.....	155
9.3.	Future Work .....	164
9.3.1.	Inclusion of Legal Frameworks.....	164
9.3.2.	Automating through an MDA architecture .....	165

9.3.3.	Integration with Security Ontology .....	165
9.3.4.	Integration with Privacy Preserving Databases.....	165
9.3.5.	Monotonic Process Changes .....	166
9.3.6.	Correlation of Privacy Rules with the Privacy Ontology .....	166
9.4.	Conclusion .....	167
10.	References .....	168

## Table of Figures

Figure 1: Real world situation .....	58
Figure 2: Real world representation .....	59
Figure 3: Privacy principles influenced by ADPO .....	61
Figure 4: External influences.....	62
Figure 5: Ontology big picture.....	63
Figure 6: Entity, Data Subject and Resource.....	64
Figure 7: Different categories of identities .....	65
Figure 8: Resources and Identities.....	66
Figure 9: Legend.....	68
Figure 10: Entity and GroupOrIndividual concepts.....	69
Figure 11: Entity and Territory.....	70
Figure 12: Entity hierarchy .....	70
Figure 13: ResourceAccessor .....	71
Figure 14: ResourceAuthoriser .....	71
Figure 15: ResourceHandler and Modifier.....	74
Figure 16: Resource, Identity and Safeguard.....	75
Figure 17: Resource and Safeguard .....	76
Figure 18: Policy and Resource .....	77
Figure 19: Policy .....	77
Figure 20: Purposes in the P3P domain .....	78
Figure 21: Privacy Process.....	79
Figure 22: ShareResource Process .....	79
Figure 23: Medical Ontology subset 1 .....	96
Figure 24: Medical Ontology - staff concepts .....	97
Figure 25: Admission process.....	98
Figure 26: EntityPerson.....	100
Figure 27: AliveDataSubject_Patient .....	101
Figure 28: Admission_ShareResourceProcess .....	103

Figure 29: B2C part overview .....	115
Figure 30: Goods hierarchy .....	116
Figure 31: Merchant trading as Business .....	117
Figure 32: Order and Delivery .....	117
Figure 33: Customer and UserAccount .....	119
Figure 34: Order concept .....	120
Figure 35: PaymentMethod .....	120
Figure 36: Customer_DataSubject and one of its identities .....	124
Figure 37: BusinessResourceReader .....	125
Figure 38: Policies and statements .....	125
Figure 39: Policy statement template in the e-commerce domain .....	126
Figure 40: Entity concepts .....	142
Figure 41: Identities hierarchy .....	143
Figure 42: Example: Identities usage .....	143
Figure 43: "identifies" object property .....	144
Figure 44: ResourceElements hierarchy .....	145
Figure 45: Resource usage .....	145
Figure 46: Statement concept .....	146
Figure 47: Privacy Process .....	147
Figure 48: Privacy principles / quality aspects .....	148
Figure 49: SecurityCriteria concepts .....	149
Figure 50: Associate class .....	149
Figure 51: Weight partition .....	151
Figure 52: "hasWeightValue" datatype property .....	151
Figure 53: Quality Assessment Criteria Value Partition .....	152
Figure 54: hasStars datatype property .....	153
Figure 55: Class hierarchy .....	154

## Acknowledgement

Foremost and all, I would like to thank my supervisor Prof. Tharam Dillon for allowing me to do this kind of research and giving me the freedom to head into the direction I chose to go. Furthermore, I would like to thank him for his continuous efforts to patiently discuss problems and issues throughout the candidature, helping me to progress towards my goals, pointing me in the right direction at times and opening my eyes to my domestic blindness. I would also like to thank him for enabling me to study in Australia through scholarships and providing me with an enjoyable and safe study and work place.

My thanks and appreciation go to Prof. Elizabeth Chang as well, who took me up at her great research institute (DEBII) and provided me with support of every kind.

Personally, I would like to thank my family, especially my parents and grandparents for allowing me to follow the way I wanted and supported all of my decisions in my life. Without their help, assistance and reassurance I would not have been able to do what I like to do and study and pursue a PhD in Australia. It

## Summary of thesis

Privacy is becoming increasingly important due to the advent of e-commerce, but is equally important in other application domains. Domain applications frequently require customers to divulge many personal details about themselves that must be protected carefully in accordance with privacy principles and regulations. Here, we define a privacy ontology to support the provision of privacy and help derive the level of privacy associated with transactions and applications. The privacy ontology provides a framework for developers and service providers to guide and benchmark their applications and systems with regards to the concepts of privacy and the levels and dimensions experienced. Furthermore, it supports users or data subjects with the ability to describe their own privacy requirements and measure them when dealing with other parties that process personal information. The ontology developed captures the knowledge of the domain of privacy and its quality aspects, dimensions and assessment criteria. It is composed of a core ontology, which we call generic privacy ontology and application domain specific extensions, which commit to some of application domain concepts, properties and relationships as well as all of the generic privacy ontology ones. This allows for an evaluation of privacy dimensions in different application domains and we present case studies for two different application domains, namely a restricted B2C e-commerce scenario as well as a restricted hospital scenario from the medical domain.

## 1. Introduction

### 1.1. Broad overview of privacy

Over the last decade, privacy has attracted more and more attention, both in the real and digital world, due to a high number of incidents relating to this issue. It is not uncommon to hear about privacy breaches on a weekly basis and they are not limited just to the academic area. Usually, the incidents reported attract high media attention and we believe that the number of unreported or unknown cases are much greater - by far. This leads one to ask, why we have some many breaches and how they could have occurred in the first place. The problem is fairly widespread and open and it is necessary to look at the fundamentals of the concept of privacy first. Commonly, we understand privacy as being the protection of "our" personal information from any kind of "misuse", that is from use in any way we did not intend it to be used. In the old days and before the invention of computers, privacy was not regarded as such a big issue as it was comparatively easy to keep track of personal details provided to other parties. For instance, if one were to buy groceries from a local shop, one would simply go there, pay for it and the transaction would be complete. The shopkeeper might remember ones face and the transaction for a while, but eventually would forget it and continues with his or her business. Nowadays, everything is recorded and once entered into a digital system, it remains there for an undetermined timeframe in many cases - regardless of one's knowledge of this, or wishes. In the case of shopping, one could try to go to a different shop every time to avoid being tracked, but this is neither convenient nor practical as there will not be enough shops to enable one to continue to do so. Nowadays, it is also common to pay cashless and shopping at different stores becomes pointless as some or all

transactions may be stored in a central location. This is especially true when signing up for shopping rewards and loyalty schemes as offered by many stores. By signing up for one of these programs, a perfect profile of shopping habits can be aggregated. Providers of such programs usually "assure" that it is used for marketing purposes only and that it will not be misused, but they usually fail to specify these terms in detail. Furthermore, one usually has no clue about the amount of information collected and stored, or its retention period, partners with whom the data is shared and so on. In most cases, it is also impossible to check the information or have it deleted - this may, however, vary according to local laws and regulations. This brings us to another important aspect of privacy: laws and regulations. They play an important role in mandating our personal information and providing some power to the subject of the information to control it. However, this will need to be examined more thoroughly at a later stage, especially in post 9/11 times when governments around the globe introduced new legislation to allow intrusion into privacy under the cover of counter-terrorism measures.

## 1.2. Examples

The causes of breaches of privacy are numerous and occur on different levels. For example, they may be as simple as forgetting to lock the door to a filing cabinet, thereby allowing someone to take a look, or simply eavesdropping on someone else's conversation. We can classify these causes in broader categories and we will show some examples of what can go wrong. Similar to the Identity Theft Resource Center, we use four to five broad categories that contain causes for privacy breaches [1]. The first one refers to loss or accidental exposure of personal information and is very common as people become aware of it quickly. In some countries,

companies are required to reveal any loss or theft of information publicly, which is the main source. The second case is insider theft. In that case, data is used differently from its consented purpose by authorized people. The third category is well known and publicized under the term "hacking", which refers to any activity of obtaining data without authorization, for example stealing or computer attacks. The fourth case is data sharing and or aggregation without consent or without necessary sanitizing actions performed on the data, which includes anonymization. The fifth category which may be seen as an umbrella, or to some extent a category that shares some of the properties of the previous ones, is called "system failure". We have experienced an example of system failure recently. The Department of Planning and Infrastructure (DPI) of Western Australia has some management rules in place that determine how staff working for the department can access, use and share their collected data. However, the guidelines provided to staff members are not entirely clear and therefore open to interpretation by the individual, which caused a worker to share information with a third party without being authorized, even though he thought he would be, and he was instructed to do so. The demarcations between categories are not clear, however, as this instance may also be considered as Insider Theft.

In a different incident that occurred between 2002 and 2003, JetBlue Airways, a carrier in the United States, gave five million customers' travel records to a US department contractor [2]. This may not sound as something terribly bad at first glance, but it violates customer privacy on a high level and sparked outrage amongst passengers and press. Essentially, the airline provided Torch Concepts, a contractor for the Department of Defense, with five million records of their passengers' travel, which directly violates JetBlue Airways' own privacy policy. Torch Concepts then acquired

additional demographic information from a different company, started merging both data sets and performed data mining to identify possible terrorist suspects. This case became public only due to the "foolishness" of Torch Concepts who discussed it publicly [3]. Nevertheless, this example shows us that data that is given away for "research purposes", which we usually regard as something "good" or something "useful" may actually be a violation of privacy. Even if we were to assume that JetBlue Airways wanted to assist Torch Concept in good faith with its research and did not receive any money or other benefits for the data sets, one may acknowledge that this may have been an unintentional and unfortunate incident from JetBlue's side that should have been checked properly beforehand. However, nobody at JetBlue seemed to have checked or adhered to their own privacy policies or notified the customers about the incident. However, it is usually the case that a company shares its data in return for some sort of benefit even if this is not immediately apparent; realistically, we assume that this may have also been the case with Jet Blue. Therefore, we classify this case under data sharing without consent or sanitization. Eventually, this incident resulted in legal action against the airline and a class action lawsuit. Any firm involved in privacy violations can experience a severe financial impact due to the costs associated with impact determination, notification and recovery as well as loss of market value and capitalization.

In E-Commerce, customer concerns about privacy protection are an increasing deterrent to the transference from the traditional means of commercial interaction to its electronic counterpart. According to [1], the number of breaches is still on the rise compared to previous years, which is frequently caused by e-commerce businesses and related activities. For example, in order to complete an online transaction, customers have to

submit their personal information such as name and address as well as their financial details, such as credit card numbers or bank account numbers. These details are then stored with the online business in a more or less secure manner. Different types of hacking activities account for the overall number of privacy breaches and identity theft. They can include weak passwords, weak or non-existent encryption of communication channels or storage facilities, phishing activities and so on. On one occasion, DSW Shoe Warehouse admitted that more than 1.4 million credit card records had been stolen from the company's database by online thieves [4] and added that personal information of about 100,000 customers had been lost as well. A continuously updated report suggest that more than 250 million records of personal information have been involved in privacy breaches since January 2005 in the United States [5]. In a recent case, the University of Alabama notified their student and staff that tens of thousands of records containing Social Security numbers had been obtained by "hackers" [6].

It is easy to see that none of the categories are crystal clear and completely independent of each other, but overlap to some extent. Nevertheless, all of the incidents can usually be attributed mainly to one of the categories, even though not necessarily exclusively. For instance, one might see successful hacking as being purely the hacker's fault, while others might see it as fault of the hacked company which might not have implemented better safeguards.

From the examples described above, we conclude that privacy goes beyond being just a matter of security, and they should not be confused with each other. We regard security as a tool to support privacy only, and should not try to achieve privacy protection through security mechanisms alone. Many

other aspects play an important role, such as policies, management and implementations as well as legislation and regulatory frameworks. It is necessary to derive knowledge from all the different aspects and areas together in order to achieve a better understanding of privacy in general as well as in the context of a certain system or particular situation, especially with regard to its protection.

### 1.3. Motivation

As we have seen in the previous section, privacy is a vital component of every individual's life, but privacy breaches are more common than ever. Naturally, it is in the vital interest of both the individual, about whom personal information is stored, as well as that of the receiving and processing entity, to protect personal information adequately. The financial cost of privacy breaches can be significant for both parties. [7] have estimated that a loss of 1000 records of personal information could cost \$US 166.272 on average. This includes costs for internal investigation, notification and crisis management as well as regulatory and compliance fees. The cost for the individual is difficult to estimate, as it highly depends on the type of information obtained. However, it can be anything from an inconvenience to thousands of dollars. The non-financial impact is even harder to estimate, but we can safely assume that a bad reputation will not be good for business.

### 1.4. Privacy by Design

When information technology was in its infancy, it was a playground for the developers and explorers. New ideas were tested and nobody thought about malicious intentions at that stage. Eventually, people became aware

that security mechanisms were needed for all the new technologies and developers started thinking about how to add them to existing technology. We experience the same problem nowadays where new tools and applications are developed on a massive scale and people are aware of security implications and applications. However, privacy is still not being considered as part of the fundamental design of those developments. Privacy is regarded as something that can be added later on top of the new exciting tools and gadgets. As we learn, we need to understand that privacy, just like security, needs to be part of the design as any later additions may contain loopholes, and privacy protection is required from the ground up.

The term "Privacy by Design" was developed in the 1990s by the current Privacy Commissioner of Canada, Mrs. Ann Cavoukian and is still promoted today [8] as it has not been addressed properly so far. Amongst others, it encompasses the need to recognize that privacy concerns and interest do actually need to be addressed from the outset. Furthermore, it requires an "early mitigation of privacy concerns when developing information technologies and systems across the entire information life-cycle" as well as the "adoption and integration of privacy-enhancing technologies" [8].

### 1.5. Scope of problem

As described above, privacy breaches can occur due to different circumstances and each of them requires different handling methods. However, as privacy breaches should be avoided in the first place, it is necessary to look at the complete picture and identify how the concepts are connected to each other. This includes the different actors and their intentions, support and safeguard mechanisms, regulatory background and

legislative issues. Furthermore, it requires one to look at the various processes and policies involved. Here it is important to distinguish between the individual - whom we will henceforth refer to as the "data subject" - about whom personal information is collected, stored, shared and processed or used in any other way and the entities using, accessing or handling it. Hence, the focus must be on the empowerment of the data subjects as it is in their vital interest to keep personal information disclosure to a limit. On the other hand, the entities using personal information need to be aware of all the implications of their use, not just in the interest of the data subject or in line with agreements made with the data subject, but also to comply with legislative and regulatory requirements.

It is not our intention to invent new privacy enhancing technologies for every situation, or create domain dependent mechanisms to enhance and protect privacy, as this would change over time with the evolution of technologies and shift in paradigms - like the one from traditional WWW to Web 2.0. Therefore, we focus this work on the creation of an application domain independent conceptualisation of the principles and mechanisms of privacy on an abstract level. Thus, the primary focus of this work is the derivation and representation of knowledge from the different concepts, processes and entities involved in the context of privacy. Naturally, many techniques exist to support privacy, including safeguards like encryption, policies and agreements, and legal requirements. Due to the vast number of elements involved, we will not provide a semantic integration of legislative or regulatory frameworks as they vary with different entities. Our system can then be used by a variety of different interest groups to enhance their understanding of the different concepts and mechanisms relating to privacy, and their implications. By using it in the design phase, it

can help to integrate privacy concepts into new systems from the ground up and helps inexperienced developers in the field. Furthermore, it can be used as a benchmarking or even certification tool by a variety of different technology providers or users to help them understand how privacy is implemented and used in their system or if, and where, flaws may exist.

As it is also not possible to include every single available safeguard mechanism or semantic concept from different areas, we limit ourselves to the most important and interesting examples that can be transposed to other techniques. This paves the way for others to develop specializations of these representations for their own purposes. Furthermore, with the continuous integration and support of other semantic conceptualizations and adoptions, our privacy "system" has the potential to evolve, capturing new concepts in more detail.

## 1.6. Plan of thesis

Following this introduction to the notion of privacy and the scope of this thesis, we will present a comprehensive literature review of the privacy domain, starting with its historical background. It is then followed by elaborations of current issues, challenges and applications, bringing the focus on to the web and semantic web respectively. In Chapter 3, we will then define the problem, describing the terms and concepts used and our choice of methodology in comparison with other methodologies available. Chapter 4 provides an overview of the solution and a roadmap to follow, while Chapter 5 describes the generic ontology that has been developed in great detail. Chapters 6 and 7 follow with an elaboration of a specialization for the domain of health care and e-commerce respectively, which are used as case studies to verify our findings. In Chapter 8, we discuss the ontology

on an implementational level and show the tools used during that process.. This work is concluded in Chapter 9 with a summary of the work, suggestions for future improvements and enhancements, and a discussion proposing the integration of this work with existing ontologies from other domains so that it can evolve and be applied to other areas.

In this chapter, we have introduced the field of privacy in general and discussed some examples of privacy breaches and violations. This was followed by an elaboration of the motivation for this work and a principal discussion of Privacy by Design. The scope of this work was then described, and the chapter concluded with a brief plan and outline of the thesis.

## 2. Review and Evaluation of literature

This chapter provides a comprehensive overview of literature relating to the area of privacy: firstly from a historical and evolutionary point of view and secondly describing the different notions and conceptions. It continues with issues of legislation and regulatory requirements, followed by a description of general problems, challenges and solutions to privacy issues, and continues our focus in digital environments like the web and semantic web. Assurance methods and identity management play an important role in privacy preservation and are therefore considered. This chapter then concludes with background information on ontologies and their relevance.

### 2.1. Privacy - A Historical View and its Evolution

The concept of privacy is diverse, has many notions and is indeed an endogenous conception. It has evolved over time and the first popular historical records regarding privacy came from Warren and Brandeis' "The Right To Privacy" which appeared in an 1890 law publication [9]. Privacy was defined as "the right to be left alone", which is a rather simplistic and outdated definition given our contemporary understanding of privacy. However, it shows that the concept of privacy was introduced under a legal constitution and hence requirement. It took almost 80 years for the concept to be revisited and revised by Alan Westin [10], who re-defined the notion of privacy as "the desire of people to choose freely under what circumstances and to what extent they will expose themselves, their attitudes and their behavior to others". This already shows how difficult it is to give an exact definition of privacy, especially since and understanding of privacy and what it entails evolves over time. We will mention a few

more recent definitions, which are not exhaustive by any means, but which show that privacy truly is understood differently by different people.

As this work's main aim relates to the area of computing, we will focus on privacy accordingly.

## 2.2. Different Notions of Privacy

Privacy can be seen in different ways and from different aspects and for each of these different aspects, different techniques for ensuring "protection" may be applicable.

The first notion of privacy comes from one of its definitions. As described previously, privacy can be defined as "the subjective condition a person experiences when two factors are in place. First, he or she must have the power to control information about him- or herself. Second, he or she must exercise that control consistent with his or her interests and values". If we examine traditional computer systems, access control is usually set up by a system/database (or whatever) administrator and usually defines which entity is allowed to use which resource. In order to enforce this mechanism, the system has to identify the entity accessing it, ensuring that it is the one it claims to be (authentication) and deciding whether or not access to a certain resource will be allowed (authorization). These rules of authentication and authorization must be set up, prior to its usage by a certain entity which usually controls the systems either partly or in full. The way in which such rules are established and enforced are governed by certain security mechanisms, such as encryption, password authentication, and so on. The entity accessing a system is usually not able to alter the

permissions to access the resources, but is limited to the rules set up by other entities and enforced by the system.

This first notion of privacy deals with access control but conversely to the one just shown. An individual with the desire to protect its privacy should be able to exercise this by defining certain rules which control the access to that information. Unlike in traditional security systems, the individual creates and alters these rules and once submitted to a computer system, it (the computer system) has to obey these rules and not vice versa. Furthermore, the individual is also able to alter and revoke the permissions granted to other entities. Open systems, where entities may not be known a priori by the system can also be supported, which is traditionally not the case. However, in order to achieve this, other mechanisms like authentication, based on certificates or reputation, as well as notions of trust and risk have to be incorporated [11].

Although the concept of privacy in terms of access control from a data subject perspective is promising, the concept of auditing or access tracking is essential and is, moreover, more powerful. Once permission has been granted to access personal information, one would like to know who actually accesses this information. It is not enough just to grant the permission for potential user and user groups, as nobody might actually use it except to obtain data about entities accessing it. In this way, it is possible to revoke information or notify entities about changes to information.

Another definition is from Privacilla: "Privacy is the subjective condition a person experiences when two factors are in place; first, he or she must have the power to control information about him- or herself; second, he or

she must exercise this control consistent with his or her own interests and value” [12]. This matches well with the definition of the Privacy Commissioner of Canada, who defines privacy as: “The right to control access to one's person and information about one's self. The right to privacy means that individuals get to decide what and how much information to give up, to whom it is given, and for what uses” [13]. The emphasis given by both definitions fits well within our context as they emphasize that the user, which we will call the "data subject" throughout the document, has to have the ability to control his or her personal information in order to control his or her privacy.

### 2.3. Privacy and Legislation

Legislation and regulatory frameworks provide a great source for notions, issues and challenges of privacy in non-private environments. Depending on the geographical (and hence political) location where technology is used and personal information is collected, different privacy protection laws apply. An overview map of all countries in the world which have privacy regulations, is shown in [14]. Usually different countries have different and multiple regulations regarding to privacy and its related laws such as the "The Privacy Act of 1974" [15] from the United States or in Europe the "Directive on Data Protection of individuals with regard to the processing and personal data and the free movement of such data" or the "European Union Directive concerning the processing of personal data and the protection of privacy in the electronic communications sector" [16]. Specific legislation relating to privacy can be found on the government web pages of the related countries. The Privacy Act of 1974 deals with privacy of information and sets some fair practices for the usage of data, which are

"transparency, individual participation, collection and use limitations, reasonable security and accountability".

Another examples of privacy legislation is the Emergency System 911 in the United States (or 112 in Europe, 000 in Australia). The Federal Communications Commission (FCC) requires all operators to obtain location information not just from landlines but also from mobile phones "within 50 to 100 meters accuracy in most cases" [17]. This also leads to some issues in the Internet world and Voice over IP (VoIP), as this system is designed to be used anywhere in the world with the same phone number. The question is, to which actual number will an emergency call be routed and what will the responder see as the number of the originator as required by the FCC?

Apart from websites, government and companies always use personal information about their customers and citizens in order to work with them. Internally, the data is usually stored in systems, which should allow access to authorized persons only. While this enforcement of access to personal data is technically possible [18], it is sometimes regarded as an overhead and increases the costs of implementing and maintaining such a system. However, a system administrator could still be able to get access to the raw data either by accessing the storage system directly if the data is not secured or by intercepting data while in insecure transit. Furthermore, the original purpose of the data usually does not matter anymore once access to personal information has been granted and it could be used in any way. Legislation and Regulation are essential for privacy preservation as it is impossible and often impractical to protect privacy by technological means. For example, as far as we know, no technology can exist that can prevent someone who has read a document, from revealing its contents to others.

However, the user may have signed a non disclosure statement and is therefore required to keep its contents confidential. In case of a breach, legislative action can be enforced if in place, which applies to many countries around the world. Another example would be the illegal distribution of music or movies on the internet. Even though mechanisms like Digital Rights Management (DRM) have been created, their success is limited. It is the legislator who can follow up on breaches regardless of failing technologies.

Hence, legislation plays an important role for privacy as the subject, depending on the laws protecting personal information, receives different protection levels. For example, personal information could be much more strongly protected (by law) in countries within the European Union (EU), than in Middle-East (ME) countries. Therefore, the level of privacy the data subject receives would be high in the EU and lower in ME countries. Hence, the data subject must be informed where and in which country the data flows and how personal information is protected before entering any data. This allows the subject to determine whether a certain service is worth the possible loss of privacy.

#### 2.4. Privacy issues and challenges on the web

As this work is mainly (but not exclusively) concerned with privacy in a digital environment, it is necessary to look at the specific issues in that context. We therefore take a closer look at current trends and issues of privacy in the digital domain. Before the creation of the internet in general and the world wide web in particular, digital storage systems existed but were not necessarily interconnected. This aided privacy as data could not be linked together as easily. With the uptake of the world wide web,

privacy became more and more an issue as it became easier from day to day to collect, store, distribute and share information. Therefore, privacy on the web faces two major problems, described by [19]: firstly, "the inherent open, nondeterministic nature of the web"; and secondly, the "complex, leakage-prone information flow of many web-based transactions that involve transfer of sensitive, personal information". Here, personal information is classified in three broad categories: personal data, digital behaviour and communication, which includes messages posted to public boards, surveys and polls. Sources of privacy violation come from unauthorized information transfer, weak security and data magnets [20], which are "techniques and tools that any party can use to collect personal data". These include online registrations, IP address identification, cookies, trojan horses, web beacons as well as federated identifies [21]. Weak security is rather common on the internet due to hacking activities or simply insecure design, or in many cases, faulty implementations of software or protection systems. Hacking activities can include the penetration of personal computers of the data subjects or other systems where personal information of data subjects is stored (e.g. shopping websites). One of the examples is the "loss" of credit card information of more than 1.4 million customers from the company "DSW Shoe Warehouse", by unknown data thefts [4]. In recent times, these "losses" seem to have increased significantly and we can find such incidents on a weekly basis; however, many of them are not published for obvious reasons.

## 2.5. Privacy preserving techniques

Various techniques to preserve privacy have been proposed and have emerged in recent times. In its initial stages, privacy preservation was the

rather simple application of security techniques to implement access control mechanisms for personal information.

Solutions to privacy protection in general and on the web in particular are based on three broad categories: protective technologies (e.g. safeguards), legislative support and social awareness. When the web (or really any other new or emerging technology) first began,, none of these was addressed properly as they may not have been known or people were unaware of them or the risks involved. Safeguarding technologies are usually thrown into the game first as they are quickly developed and released. The other two take their time and are mostly reactive and not proactive - until something happens. Technological solutions to privacy concerns on the web are plentiful and have evolved over time. Most of them are based on traditional security technologies, like encryption. A taxonomy of technology- and regulation-enabled solutions for privacy preservation can be found in [19]. A main driver for privacy on the web is the "Platform for Privacy Preferences (P3P)" [22] and its related "A P3P Preference Exchange Language (APPEL)" [23]. P3P allows websites to define privacy policies for information from the users that is used by that website. In turn, the users have the ability to decide whether to use the website and transmit their information by defining their preferences through APPEL. This process of evaluation and comparison can be semi-automated as policies are written in standard XML. In most cases, the web browser takes responsibility for that action. We consider P3P as a static facility that has a defined number of concepts and provides a description of the various implications of using them. It does not take into account the processes and multiple and consecutive interactions, and also does not refer to the different privacy dimensions and their implications for the overall privacy experience. It is also not possible to apply privacy preferences from one application domain

to a different one as they are described in an application domain-specific manner. Furthermore, we have to note that it is not possible to specify what 'acceptable' is, but only what NOT acceptable is. This has partly been addressed by other projects like "XPref" [24], an XPath-based Preference Language for P3P.

In our earlier work, we have described a privacy management system for mobile devices in context-aware environments [25]. However, this system was limited in itself as well, as it could not be abstracted to other application domains and did not specify or classify the dimension of privacy precisely.

## 2.6. Privacy issues and challenges on the semantic web

The semantic web is (still) a vision of a global network where software agents can communicate with, process and understand each other and the meaning of data. This is achieved by annotating documents and linking information together - for example, by means of ontologies. Ontologies are "a shared conceptualization of a domain on which all parties agree" [26]. A different and more explicit definition comes from Guarino et al [27], who see ontology as "a formal shared, explicit, but partial specification of the commonly agreed upon intended meaning of a conceptualisation" [28]. If software agents facilitate such an ontology, they can exchange information, interact with each other and derive information that is not explicitly given. For example, a software agent could determine that the passport of its user is about to expire, locate the facility for renewal, determine the time required and (perhaps in the future) request it automatically or notify the user to get it done. It is likely that personal information may need to be transmitted, stored or located in these transactions and privacy is of

importance. Early attempts to govern access to web pages, which have been enhanced with annotational markups have been very simple and did not make use of the shared knowledge - hence, it was not considered as facilitating an ontology.

In recent times, the research community has acknowledged the need for and usefulness of ontologies on the semantic web [28-31], especially to build a usable, useful and pragmatic semantic web, not just academic one. Ontologies have been built and proven to be quite successful, like Gene Ontology (GO) [32], Protein Ontology [33] or Trust Ontology [11] and many more - all representing a piece of the real world as understood and used by experts in that particular domain. Therefore, it comes as no surprise that the concept of privacy may be modeled as an ontology to aide in the preservation of personal information. Semantic web concepts and techniques have been used in recent times to support security and privacy. We reuse some of those concepts in our ontology as they are building blocks, in particular the security ontology mentioned below.

#### 2.6.1. Semantic and ontological concepts

As a first step to support privacy, mechanisms for access control have been proposed, which were very similar to the traditional ones and did not make proper use of semantic concepts. Denker et al proposed "Access Control and Data Integrity for DAML+OIL and DAML-S" [34]. Therefore, the access control mechanism itself is based upon an ontology, allowing extensions by future implementations. However, the ontology is used in a rather simplistic way, as it is not related to the actual semantics of the content. Tumer et al. describe a semantic-based privacy framework for web services [35]. Basically, it allows users to specify their privacy preferences with

different permission levels. However, its concepts and privacy mechanisms are tied to the domain (web services) and cannot be used independently. A similar but more sophisticated approach has been proposed by Qin et al. [36]. They include ontological mechanisms to derive access control on a conceptual level. For example, it is possible to determine that a missile is a weapon and because access to weapons is denied, access to missiles must be denied as well. However it refers to security concepts only and does not take other dimensions into account.

Raskin et al. were among the first to introduce the concept of ontologies to the domain of information security and their proposal included: 1) natural language data sources as an integral part of overall data sources in information security applications; and, 2) formal specification of the information security community know-how for the support of routine and time-efficient measures to prevent and counteract computer attacks. It includes a comprehensive list of words and phrases which were related to the domain of information security at that time. All of these concepts are modeled item by item within the ontology.

Based upon Tropos [37], a Requirements Driven Development Methodology, Mouratidis et al. extend the approach in order to incorporate security into the model [38]. Tropos itself is a rather complex and complicated approach and requires an overall change in the way programmers think as it introduces the concepts of actors, goals, soft goals, tasks, resources and social dependencies. They are intended to aide programmers to model the real world more efficiently and with less proneness to error. However, due to the major changes involved in programmers' thinking, it may not be sensible to adopt it as such, even to incorporate and evaluate security in application design and programming.

Many of the previous approaches have in common that, although they try to preserve privacy, they mostly use security safeguards as their only approach which may be considered as security protection and not privacy preservation approaches. Schumacher, on the other hand, created the first proper security ontology. It is an extensible core security ontology that incorporates the concepts of asset, stakeholder, security objective, threat, attacker, vulnerability, countermeasure and risk [39]. The ontology is based on security (design) patterns, which are concepts to help improve the understanding of security. According to their specification, a "security pattern describes a particular recurring security problem that arises in a specific security context and presents a well-proven generic scheme for a security solution". These patterns are closely related to Object Oriented (OO) design patterns, which are used in software engineering [40]. The ontology is abstract and not specific to any particular language and prevents inconsistencies in the terminology. This security ontology can be easily reused by other ontologies, for example by a privacy ontology, which can use it to provide and evaluate safeguards for the protection of personal information on a high level of abstraction.

In their paper "Towards Cross-Domain Security Properties Supported by Ontologies" [41], Sure and Haller present an approach for secure and trusted collaboration between different businesses. It is mainly based upon secure authentication of the counterpart (e.g. the other business) facilitating a public key infrastructure. The core concept of this security ontology is the X509 certificate standard enriched with different properties and rules. By using a semantic layer (the ontology), it is possible to span administrative domains and establish trust between them.

The authors of "Authorization and Privacy for Semantic Web Services" [42] address security of semantic web services in a declarative way. The descriptions of semantic web services are in OWL-S and related ontologies created in order to annotate them. They include security characteristics and mechanisms like encryption and digital signatures and a proposal is made to incorporate privacy and authentication policies into these OWL-S descriptions. Policies related to aspects of security, privacy and authorisations are expressed in the Rei ontology [43], an RDFS-based language for policy specification. However, as policies and security are some of the dimensions of privacy only, this approach is fairly limited in terms of expressiveness and its overall understanding of privacy.

## 2.7. Privacy preserving techniques in Data Mining

Although we do not intend to propose new techniques for privacy preservation in data mining, we will use existing and upcoming techniques as safeguards within our ontology. It is therefore interesting to investigate this area, among many others, in order to understand the principles involved and how they are used to achieve better privacy preservation.

Privacy preservation in data mining is one of the biggest challenges today as data mining concepts and algorithms get better over time to try to discover "hidden" knowledge in large datasets. Randomization was introduced first to add noise to a data set and make it less accurate and hence less prone to attacks [44, 45]. One of the earlier of these approaches includes a technique called k-Anonymity [46], [47], [48-53], which was developed to minimize the impact of (re-)identification through queries on public databases. Essentially, it is a method used to alter the granularity of a query and output results only in cases of k or more results as it may be

possible to infer the actual identities of the results and lose anonymity. The l-diversity model was introduced [54] afterwards to deal with some of the weaknesses of k-Anonymity, where the lack of diversity in the k-anonymized dataset leaves it open to strong attacks. Nevertheless, neither approach scales very well in the case of high-dimensional databases [48, 55]. More recent research in the area led to the expression of "Differential Privacy", which "intuitively, captures the increased risk to one's privacy incurred by participating in a database" [56, 57]. With the assistance of these techniques, it is possible to obtain accurate information about a database while providing a high level of privacy at the same time as they provide "formal and ad omnia privacy guarantees" that are proven rigorously.

## 2.8. Summary

In our literature review, we described the different definitions, understandings and notions of privacy. We noted that many seem to mistake security for privacy and do not address or classify the various dimensions of privacy. It seems that - to date - there is no coherent semantic representation of privacy available that could aid in the understanding of this endogenous concept. Furthermore, none of the technologies discussed can be used to assess or benchmark the level of privacy experienced in certain situations or systems as no clear and precise criteria are defined that could address this issue. Finally, it seems that is not possible to use the various methods and techniques elaborated easily across different application domains by extension.

### 3. Problem definition

In this chapter, we present a clear and concise definition of the problem addressed in this thesis. Nevertheless, before we can begin the elaboration, we define terms and concepts and our use of them. Afterwards, we discuss our choice of methodology to address the problem as stated in the problem definition.

#### 3.1. Terms and concepts used

As this work is mainly concerned with user perceived privacy, we are bound to present our interpretation and usage of this term. Therefore, it is important to understand that our focus is on privacy in the private and commercial sectors, as privacy in the governmental environment is categorically different [12].

Furthermore, as the concept of privacy is closely related to the terms Trust, Reputation and Security, we also provide their definitions and usage in this thesis.

##### 3.1.1. Privacy

Privacy has many different definitions, which have evolved subsequently over time. Different definitions like "the right to be left alone" [9], "the desire of people to choose freely under what circumstances and to what extent they will expose themselves, their attitudes and their behaviour to others" [10] or "freedom from unauthorized intrusion" [58] show the difficulty in defining the concept of privacy. There are several important dimensions to privacy and they

include access control, the right data and the right purpose as discussed below.

#### 3.1.1.1. The right data

A further notion of privacy deals with the concept of the right data which is usually seen as information being accurate, complete, relevant and timely [59]. It is strongly desirable for the subject of that information, that it actually is the right data in terms just described. Imagine that someone alters personal information in order to gain advantage for him- or herself. For example, if two people apply for a position and it is possible for one of them (i.e. the boss of the other) to alter personal information of the other one in order to get the job, this would be regarded as a loss of privacy. Not just because the person may have made the change illegally (no permission to access it), but also if it was done legally (the person having access to that data), a violation of privacy occurs in either way as the purpose would have been compromised. The subject of that information should be the only one to make those changes, either directly or indirectly with his or her consent.

#### 3.1.1.2. The right purpose

Once personal information about oneself has been submitted to a system, one has lost direct control, and hence, control of privacy becomes both more necessary and difficult at the same time. The notion of the right purpose concerns the information and its usage [59]. Unlike the previous notion of access control, one might have granted someone access to certain personal information. However,

one cannot be sure what the other entity does with the data. When one submits data, it is usually done for a certain purpose, mostly to gain something (e.g. buy a product, obtain information about something, etc.), which is then regarded as worth the loss of privacy or mandatory to fulfil the purpose (e.g. the postal agency needs a name and address to actually deliver the goods). One is usually informed (explicitly or implicitly) about the usage of the personal information by the other party, or as in real life, it is implicit. However, generally, no mechanism exists that could stop the other party (which is authorized to access it for the sake of simplicity) from accessing that data and use it for purposes other than the one explicitly or implicitly agreed on.

It is easy to understand that all the notions just mentioned are part of the big picture of understanding privacy and creating a framework that includes those properties.

#### 3.1.1.3. What about Confidentiality?

Confidentiality, which is regarded as a property of security is not the same as privacy. Confidentiality deals with the protection of information in order to ensure that only authorized people get access to what they have been authorized for. Young states that "confidentiality is typically desired where people do not act in their private capacities and wish to protect the interest of somebody in which they have some non-private role", which is in contrast to the definitions of privacy [60].

#### 3.1.1.4. Definition of Privacy

As we require a precise enough definition, the previous ones are not sufficient enough for our purposes. It is necessary to emphasize that privacy is not about information itself, but about a data subject's power to control and release personal information. Hence, in this thesis, we define privacy in the following way:

#### Definition

"Privacy is a subjective condition a person, whom we call the data subject, experiences. Hence, the data subject must have the power to control information about him- or herself and must exercise that control consistent with his or her own interests and values".

This definition is similar in conception and conforms to the statements made by Privacilla [12] and the Privacy Commissioner of Canada [8]. In a commercial environment, it is relatively easy to **control the release** of personal information, as one can decide not to participate at all in a certain transaction and therefore withdraw from the whole process completely if deemed necessary. This refers to the existence of the legal power to control that release and is independent of its pleasantness. On the other hand, controlling that information with one's own interests and values in mind, is becoming more and more difficult due to the inherent nature of information storage and processing in a complex network of business processes. The protection of privacy is therefore to be sought mostly in this area.

#### 3.1.2. Trust & Reputation in Privacy

Trust and Reputation are closely-related concepts that play an important role when dealing with service providers. Chang et al [11]

define trust as "the belief the trusting agent has in the trusted agent's willingness and capability to deliver a mutually agreed service in a given context and in a given time slot". In our case, we are essentially concerned about the context being "privacy preservation". Given this is the case, quality aspects have not yet been defined in previous works and this thesis aims to address this by providing quality assessment criteria. In our context of privacy, this essentially means that a data subject who has dealt with a service provider previously is able to assess the level of trust he or she is willing to give the service provider with regards to the level of privacy preservation in previous dealings. On the other hand, if a data subject has not dealt with a particular service provider in the past, trust measurements cannot be applied and one has to refer to reputation. Agent reputation, which is closely related to our need to know the reputation of an entity, has been defined by [11] as "an aggregation of the recommendations from all the third-party recommendation agents, in response to the trusting agent's reputation query about the quality of the trusted agent". We are using reputation in the context of privacy and can therefore define the dimension of quality from the previous definition as "privacy protection". This means that any reputation queries are about the trusted agent's privacy protection capability as experienced by third-party recommendation agents.

### 3.1.3. Security & Safeguards in Privacy

Security is one of the more important dimensions of privacy preservation, but it is important to mention that security is different from privacy as described earlier. Security in the context of privacy usually refers to the concept of "Safeguard" that is put in place to

achieve privacy preservation on different levels. A safeguard is therefore defined as any technique that prevents unauthorized entities from gaining access to personal information. "Safeguard" however, does not refer to methods of preventing access to personal information to authorized entities for unauthorized or not consented purposes.

Different safeguard categories exist, which we define as "Transit-Safeguards", "Storage-Safeguards" and "Access-Safeguards". It is important to distinguish between the three as it is usually impossible to apply a safeguard from one category to a requirement of another one. Furthermore, it is also possible to draw conclusions from the usage of those safeguards such as the data will be transferred - by any means - to a different location or the data will remain in the same place. In many cases, it will be necessary to apply both of them, independently of each other, to certain scenarios as there is a need to collect data (which is transit), store it (non-transit) and access and process it (transit). Examples of a storage-safeguard would be encryption with a particular encryption technique, but also the physical protection of documents by a lockable filing cabinet. On the other hand, a transit-safeguard might be as simple as a (physical) envelope or using a particular tunnelling technology like Transport Layer Security (TLS). Access safeguards provide mechanisms to authenticate and authorize individuals to access personal information on various levels.

#### 3.1.4. Entity

##### Definition

The term "Entity" is defined as a "physical or digital agent, person, company or service provider that falls within a certain jurisdiction".

We are bound to define this term as broadly as possible in order to not limit our scope to any particular context and provide the highest level of abstraction. It is noteworthy however, to mention that this does not apply to governmental institutions as such due to the categorical difference between privacy in the private and in the governmental sector. In this thesis, "Entity" refers to concepts at the class level, while we define instances of this class as "entity-instance". It is important to make that distinction as we encounter both principles throughout the thesis. For example, the data subject is an entity class, while a particular data subject is an entity-instance. We are using the term "entity" in a way that differs from the way it is used in relational databases for example as it refers to agents as stated above and contains behavioural characteristics.

#### 3.1.5. Data Subject

##### Definition

The Data Subject is defined as "special case of entity who is interested in preserving a certain level of privacy upon his or her own personal information with regards to any other entity". This refers to the class level and not to a particular instance of that class.

#### 3.1.6. Resource

##### Definition

We define a resource as "any piece of information about an entity".

This definition includes personal and non-identifying information as it is necessary to look at both kinds of elements, particularly when combining them.

### 3.1.7. Personal Information

#### Definition

Personal information is defined as "a resource with the potential to identify a particular identity, both standalone and in combination with other resources".

### 3.1.8. Identity

#### Definition

We define "identity" as "an intentional accumulation of suitable personal information that together distinguish one entity from another".

The need for identities has been widely acknowledged and is a mechanism used to create anonymity and pseudo-anonymity. The latter case is often used to participate under a certain pseudonym in certain transactions consistently, without revealing any or all personal information about the underlying real world entity. This may be as simple as a different name in online chats, a different email address in social networks or a falsified certificate of birth in the real world. We make no judgements about the rights and wrongs of using identities as this refers to legal issues as well as moral and ethical ones. In order to

eliminate fake identities that may be used to mislead people, we can utilize the mechanisms of trust and reputation [11].

#### 3.1.9. Process

##### Definition

We define "process" as "any number of transactions performed by any number of entities to achieve a certain goal under certain conditions and restrictions".

This is a very generic assumption of a process, but we will provide more specific details in the context of privacy at a later stage.

#### 3.1.10. Policy

##### Definition

A policy is "a repository of statements provided by an entity with a certain purpose about a particular resource intended for a particular audience or entity, which is valid for a particular timeframe".

#### 3.1.11. Consent

##### Definition

We define consent as "the mutual agreement between two or more entities about the conditions and purpose of certain resources".

In many cases, consent is established between a data subject and another entity to define the conditions of usage of personal information for particular processes that involve the handling of

personal information. This is not just limited to usage and sharing of personal information.

### 3.1.12. Repository

#### Definition

A repository is defined as the "total collection of resources" that is manifested in persistent storage.

## 3.2. Privacy Principles

We use privacy principles and its synonym "quality aspects" throughout the thesis to assess the level of privacy as elaborated earlier. We define them in accordance with [61]. It is important to mention that they vary between territories and legal boundaries. Researchers in this field acknowledge that the following ones are essential and provide a concise representation of the knowledge of the real world.

### 3.2.1. Data Quality

The data quality principle refers to the idea that processing of personal information is to comply with quality requirements. Therefore, the quality dimensions refer to personal information that is adequate, relevant, not excessive, correct and accurate to the purpose for which it is collected and subsequently processed. This includes the terms of storage, periodic clearing, information about corrections and disclosures to other parties, measures to minimize, detect and handle errors as well as an authorization inspection for data input.

### 3.2.2. Transparency

The transparency principle refers to the idea that the data subject must be informed about what is done to his or her personal information. This includes the provision of information about the collection process before or during the collection when personal information is provided directly to the data subject; or if obtained indirectly and a notification is impossible or economically not viable, the recoding of the source of data. Furthermore, the purpose and need of the data collected and any third party involvement in its processing needs to be provided to the data subject. Finally, the legislative framework upon which the data is collected needs to be provided.

### 3.2.3. Intention and Notification

This principle refers to the conditions of any intention to collect and the subsequent notification dimensions. This includes the timeliness of notifications, nature of processing, name and address of the data controller, the categories of resources that are (potentially) collected, the recipients of any resources, descriptions of any safeguards and the purpose of the resources. However, this principle is not used throughout the thesis any further as it defines preliminary requirements that have to be met (in certain countries and regions) before collection of personal information becomes legitimate.

### 3.2.4. Finality Principle

The finality principle states that personal information is collected for certain agreed purposes and may be used for these only. However,

there may be legal requirements to make the data available to other parties under special conditions. This may be a regulation relating to national security, but can also include the data subject's vital interests, like obtaining medical information in an emergency. The finality principle also provides rules to keep data only as long as necessary for the intended purpose or apply anonymization methods otherwise.

#### 3.2.5. Legitimate Grounds of Processing

This principle refers to the idea that personal information may be collected and processed only if legitimate grounds can be found. This includes unambiguous consent from the data subject that must have been given without any pressure. Other grounds may be legal requirements or the protection of the vital interests of the data subject. It can also include reasons relating to public interests. There are some limitations to the collection that apply to specific categories that include, but are not limited to: religion, philosophical beliefs, race, political opinions, medical data, sex life, trade union memberships, data about criminal conviction and unlawful behaviour.

#### 3.2.6. Data Subject's rights

The data subject has certain rights within the legal framework with regards to accessing, processing and rectification of personal information. They include information about how data processors process personal data, who the data is shared with and how they intend to use it. It also includes the rights of the data subject to request rectification, supplementation, deletion or blockage for personal information for certain or all purposes. Hence, it is important

to set time frames for the carrying out of the changes and notifications by the collector if adjustments cannot be made for certain other processors and third parties.

#### 3.2.7. Security

The security principle requires data collectors and processors to implement appropriate technical and organizational measures to protect personal data against loss or any form of unlawful processing. Hence, it is important to emphasize that an adequate level of state-of-the-art security to be applied, depending on the type and category of data and the nature and of the risk involved during processing.

#### 3.2.8. Accountability

Any entity that is processing or handling personal information is accountable for the compliance with any of the principles listed. Therefore, the data subject can contact the entity or entities in question if its data has been processed or used not in accordance with those principles and may have legal rights to pursue action against the entity if not satisfied with the outcome.

#### 3.2.9. Openness

This principle deals with the openness of information regarding policies and procedures. They need to be ready and available at all times for inspection by the data subject and/or relevant authorities.

### 3.2.10. Anonymity

The privacy principle of anonymity refers to the idea that personal information must be transformed in such a way that identification is impossible afterwards. It may also be possible to apply pseudo-anonymity in order to identify personal information within a certain processing domain and in such a way that they can be linked together by a trusted entity only.

### 3.2.11. Transfer of personal information between different jurisdictions

Legal privacy protection varies dramatically between different jurisdictions. Therefore, it is necessary to limit transfer of personal information to countries with lower privacy protection laws unless unambiguous consent has been given by the data subject or it is necessary for the fulfillment of a contract between the data subject and the processor. Furthermore, it may be required if it is in the vital interest of the data subject or is a legal requirement.

## 3.3. Problem definition

Privacy preservation is not just a technical problem. It involves different mechanisms on different levels. It is necessary to take the entities processing and their jurisdiction into account. It is also necessary to look at the preferences of the individual users with regards to their own preference for privacy protection as it as an endogenous conception. However, it is difficult for system developers and data processors to obtain a partial, much less a complete understanding of the different concepts and dimensions of privacy as many are not experts in that domain. This is

especially true when operating on a global scale where requirements change between different countries and may fluctuate with cultures. However, it is necessary to understand privacy principles in order to adhere to them and build a system that is compliant with the different rules and regulations and is transparent enough for data subjects.

### 3.3.1. Summary of problem definition

This thesis addresses the problem of capturing the knowledge in the field of privacy, with its concepts and relationships. This is done in a way that permits (semi-)automatic processing and evaluation of privacy protection levels experienced. This helps with the user experience with regards to privacy preservation and can be used by users to apply their privacy related preferences across multiple application domains. It also assists application developers to adhere to privacy legislations in a particular regulatory environment.

The research questions that need to be tackled so that all the different aspects of this problem are addressed are therefore:

- 1) "How do we represent the various privacy concepts and relationships and the way they link up with each other?"
- 2) "How can such a representation be used to increase the level of privacy preservation for the data subject?"
- 3) "How can such a representation be used by system developers and service providers to achieve compliance with current rules and regulations as well as state of the art techniques to safeguard personal information?"
- 4) "How can this representation be used throughout various application domains and provide cross-domain privacy preservation experiences?"

We strongly acknowledge and have mentioned previously the need to involve legislation and regulatory frameworks in this work. However, as we are not experts in the legal domain, we cannot provide a semantic framework for this purpose, but will limit this work to some simple examples from the legal domain. It would be beneficial (and we will consider this in future works) to conceptualize the various legal privacy requirements on an abstract level that can be used (semi-) automatically to derive knowledge from that domain.

In order to look at the different levels and to determine privacy preservation requirements and the level of privacy experienced by individuals, it is necessary to represent the knowledge in the domain within a conceptual framework.

#### 3.4. Choice of methodology to problem solving

In this section, we describe the methods for problem solving in information systems. As privacy is such a diverse and endogenous conception that varies across different application domains, we need to solve the research issue by following a systematic scientific approach, namely science- and engineering-based research methodologies. Therefore, we will give an overview of existing methods justify our choice of methods.

One could choose several different approaches to represent this privacy knowledge and they include: 1) Natural Language Processing (NLP); 2) Mathematical representations; and, 3) Knowledge representation, which can be categorised further. This discussion is elaborated below.

##### 3.4.1. Natural language processing

Natural language processing refers to the extraction and processing of knowledge from normal written text without any particular structure or schema behind it. In the field of privacy, privacy policies as well as legal documents provide a source for such an approach. Therefore, one could assume that using NLP techniques to extract the relevant keywords and passages could yield in a determination of the level of privacy experienced. However, this does not make use of the inferred knowledge from the different privacy concepts and how these work together. Additionally, the concepts and relationships might not be exposed and easily visible. It is also very difficult to apply these to different application domains without major changes. As the documents in question are not precise either and have different terminology that might even change throughout the document, NLP seems to be insufficient to help solve the privacy preservation problem. Finally, not all sources that determine privacy implications are written in plain text and therefore they could not contribute to this process.

#### 3.4.2. Mathematical representation

A mathematical representation of privacy would be most likely be unsatisfactory. This is due to the endogenous nature of privacy and how it varies across different people, countries and application domains. It also changes over time and with changing requirements and preferences of a particular data subject, it may require new technologies and concepts that have not been created or used previously. This would make it necessary to provide new mathematical

calculations every time a new application domain is introduced or new technologies evolve that enhance the privacy preservation experience.

#### 3.4.3. Knowledge representation

Knowledge representation seems to be a good way to deal with privacy across different application domains. It usually requires experts in the domain to collect, structure and apply the knowledge appropriately and provide guidelines on how to use it. This may be done semi-automatically or collectively by a community.

#### 3.4.4. Glossary of terms

An initial idea to represent privacy and its different concepts could be a glossary of terms. This is done by an expert in the domain and lists all the different concepts. Furthermore, it provides synonyms and acronyms as found in the domain and relates them to each other. It can also describe the various attributes of concepts. However, a pure glossary of terms would not assist us in achieving our goal here as there are no taxonomies among the concepts or any kind of relations between the different concepts - except synonyms and acronyms of course. Without taxonomies and relations it is impossible to determine or even infer anything from the concepts provided. Although it is fairly easy to extend such a glossary by simply adding new concepts, it would be impossible to process them in any kind of automated way as the processing agent would by no means be able to understand them.

#### 3.4.4.1. Topic Maps

According to Wikipedia, Topic Maps is a "standard for the representation and interchange of knowledge, with an emphasis on the findability of information" [62]. They are basically used to link and associate different concepts together and provide information resources to a particular concept. While this may appear to be a viable approach, it does not seem to support our need to define the allowable types in order to categorize and evaluate privacy preserving mechanisms. This makes it hard to maintain a sufficient structure throughout the knowledge representation, especially for large bases that can be extended properly and still be used by relevant applications without modifications.

#### 3.4.4.2. Ontologies

The previous mentioned approaches are all good starting points to represent knowledge in the relevant domain. However, they lack a certain structure and associations among the concepts and attributes described. According to Guarino et al., an ontology is "a formal shared, explicit, but partial specification of the commonly agreed upon intended meaning of a conceptualisation" [27]. The ontology approach fits well with our requirements of representing knowledge in a particular domain and structuring it appropriately. Building ontologies however, can be achieved by different methodologies and the one we have chosen is METHONTOLOGY [63]. The fact that - to our knowledge - no other ontology exists in the area of privacy and METHONTOLOGY

supports building an ontology from scratch, makes it a candidate of choice. Furthermore, the METHONTOLOGY approach also allows us to incorporate knowledge during the building process in form of other ontologies from other domains. This is true especially in application domains, but also by adding upper ontologies to make it compliant and compatible with already established core and application domain-independent concepts. Finally, the METHONTOLOGY approach allows us to design the ontology in an implementation independent way, thus allowing one choose the most appropriate and perhaps standardized language or most widely accepted language. In order to develop proof of concept, we had to select an implementation language and a tool to help with the development. The tool for implementation selected was Protégé which, amongst others, generates code in the Web Ontology Language (OWL), which represents the generic privacy ontology. OWL is a standard of the W3C to define ontologies on the implementation level on the semantic web and therefore a suitable language for our implementation as it is widely accepted and standardized and based upon XML. Naturally, every language and its implementation has its own limits and restrictions and slight alterations had to be done to some ontological concepts and associations to conform to these when entering them in Protégé. However, the modifications are merely of a technical nature and would not impact on the expressiveness or the semantics of the ontology. An instance of such an alteration is the lack of support for trinary (or higher) associations between concepts in OWL, and therefore, intermediate concepts are created to support such association types.

## 4. Solution Overview / Roadmap

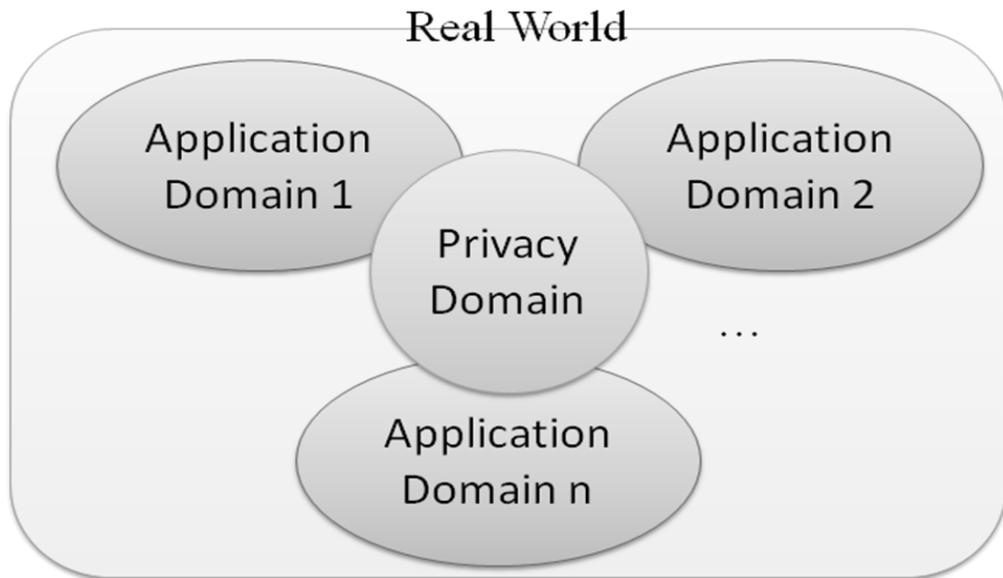
This chapter gives an overview of the solution provided by this thesis and can be seen as a roadmap for understanding the different components and their interaction. We start by showing the initial development of the ontology followed by the concepts and relations that have been identified.

### 4.1. Ontology development

As elaborated in Chapter 3, an ontology is a feasible method for capturing and structuring the knowledge in the domain of privacy. As we have opted to use METHONTOLOGY as our development methodology, the glossary of terms and relations had to be established from the domain knowledge. We have used various documents from the legal sector, including the "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" [16] and the "OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data" [64], in addition to research work carried out by the PRIME [61] and PISA [65] projects as viable sources for those concepts. Although mentioned in these documents, many concepts appear to be mostly unstructured, sprinkled throughout the documents without detailed rigorous definitions of the concepts and relationships. We clearly address this lack of precision. We facilitate the privacy principles as our quality aspect measures and include Data Quality, Transparency, Intention and Notification, Finality Principle, Legitimate Grounds of Processing, Data subject's rights, Security, Accountability, Openness, Anonymity and finally, Transfer of personal information between different jurisdictions as defined in Chapter 3 and discussed in our previous work [66].

#### 4.2. Ontology architecture

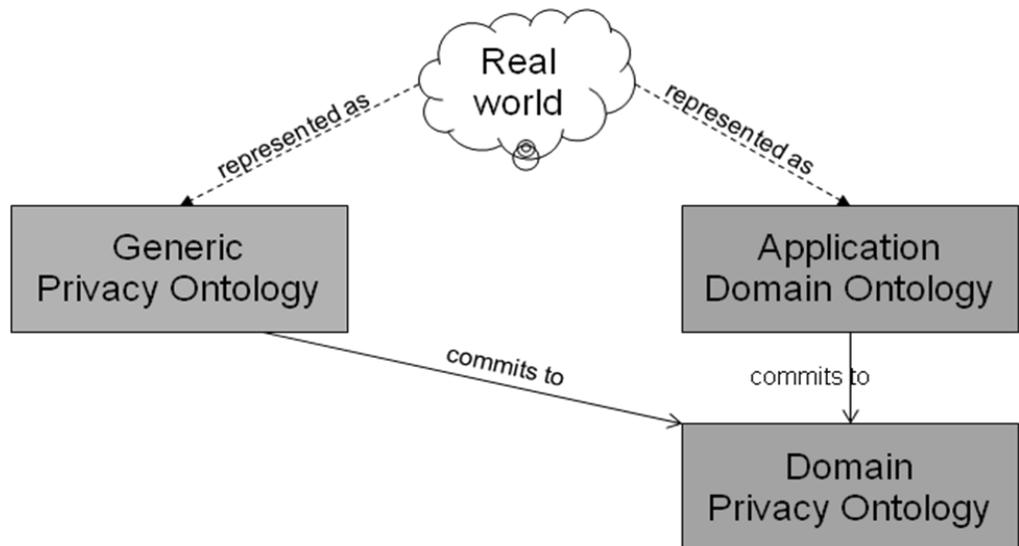
The general idea behind our approach is that there is a distinction between knowledge in the domain of privacy and its application in specific domains as shown in Figure 1.



**Figure 1: Real world situation**

The notion of privacy in a domain such as medicine or health will be somewhat different from that in e-business. Notwithstanding this there are certain core concepts which are common to all application domains. This leads us to propose an architecture as depicted in Figure 2 with:

1. a generic ontology which has a representation of the core concepts of privacy
2. specialized ontologies which represent commitments to the concepts of the generic ontology but which, in addition, describe specialized constraints, concepts and relationships pertinent to the application domain.



**Figure 2: Real world representation**

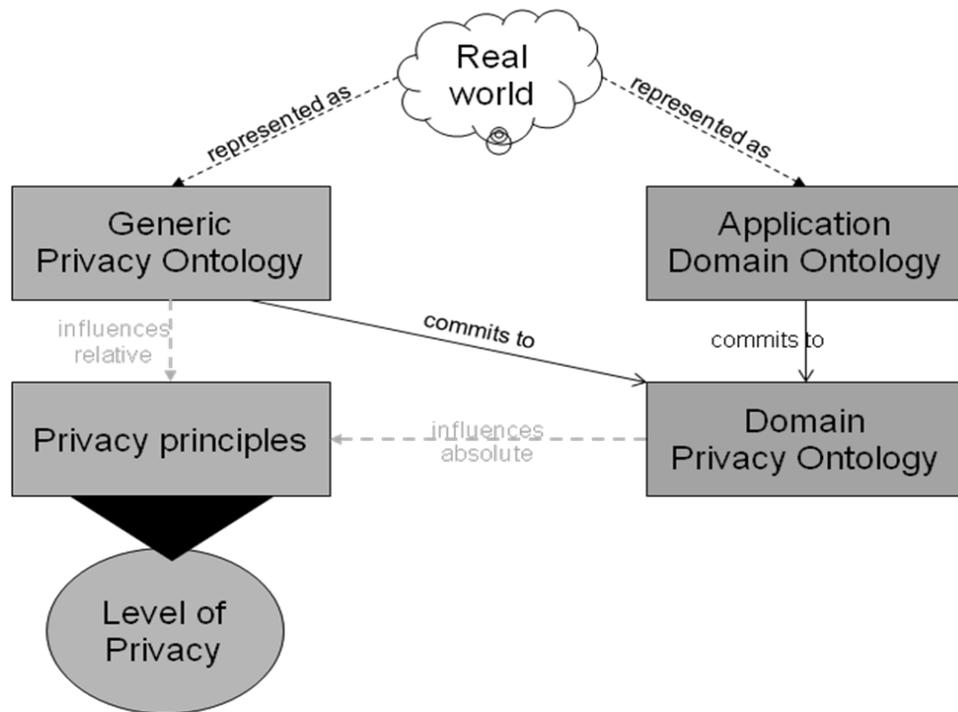
This idea of separating certain elements into a generic ontology and specialized extensions, the ontology commitment, is not new and has been successfully proposed, implemented and used by various authors, like Spyns et al. [28, 67] or Wouters et al. [68]. The approach was first proposed by [67] in an attempt to decompose an ontology into "an ontology base, which holds (multiple) intuitive conceptualisation(s) of a domain, and a layer of ontological commitments, where each commitment holds a set of domain rules". Hence, a classical database model-theoretical view is adopted. It separates the conceptual relationships from the domain rules, which are conceptually moved to the application domain. Each of the ontology commitments in turn mediates between the generic ontology and the application domain ontology.

The generic privacy ontology (GPO) with the core privacy concepts are completely application domain independent. The privacy concepts within the core privacy ontology have relationships with the privacy principles, which we commonly refer to as "quality aspects" as described in [11] and "influence" their values. As discussed above, privacy is not a

mathematically precise definition, and therefore we make use of fuzzy terminology when assigning levels of influence to the privacy principles. As the generic privacy ontology is application domain independent, the influential values assigned are not just fuzzy, but also have a relative character. A certain concept or process within the generic privacy ontology influences a particular privacy principle without actually having a need to state proper value, for example; this means that the use of that particular concept has a certain influence on the associated privacy principle, which needs to be specified more precisely in the extension. We will provide the different levels of influence, which refers to the quality assessment criteria in chapter 5. Each privacy principle has certain quality assessment criteria that are not necessarily disjoint from the assessment criteria of other principles. Therefore, the different aspects influence each other in addition to the influences experienced by the concepts of the generic privacy ontology.

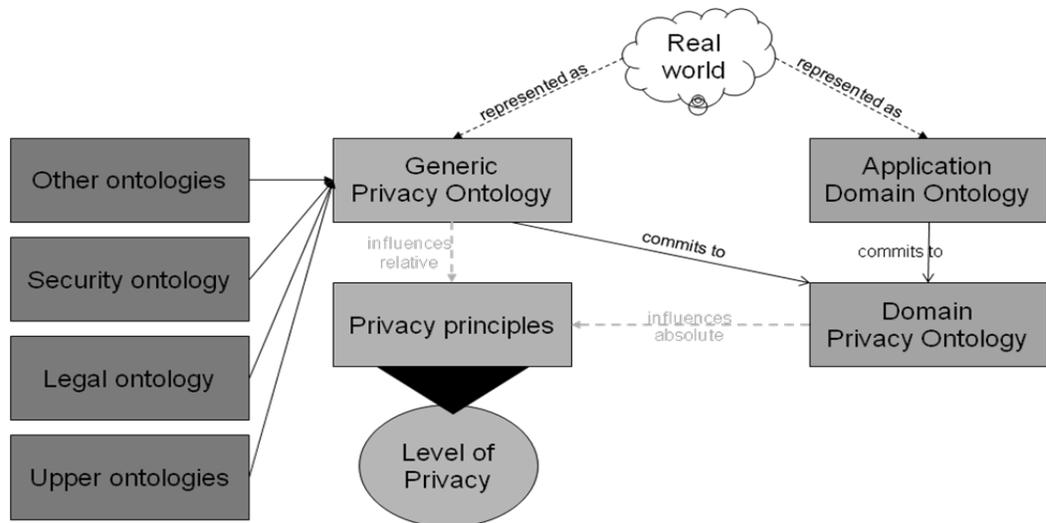
With the assistance of the core privacy ontology itself, we cannot evaluate the different levels of privacy experienced due to the relative nature of the influential levels. Therefore, the generic privacy ontology needs to be extended for a particular application domain by a domain expert. This is also known as ontology commitment as the specialized ontology commits to using all the higher-level ontology concepts and specifications. The domain expert would either use an existing ontology from the application domain or would have to create a new one in the domain if no suitable one were available. This application domain ontology (ADO) is then used to derive the concepts for the application domain privacy ontology that is based upon the generic privacy ontology. This needs to be done once during its initial creation and whenever new technologies emerge in that domain or the domain concepts evolve. A domain expert is therefore

guided by the generic privacy ontology and can assign more concrete values to the influential relationships as depicted in Figure 3. The terminology of influence used here is comprised of "very high", "high", "medium", "low", "unknown", "yes", "no" or "not stated", but can also include proper values that define a certain assessment criteria like "health data", to which a value from a scale from 1 to 5 is assigned. However, it is not necessary to use all of the quality assessment criteria levels, as a subset for each might suffice, depending on the actual principles. This is then described on a star scale, where we assign five stars to a level of "very high" and one star to "unknown".



**Figure 3: Privacy principles influenced by ADPO**

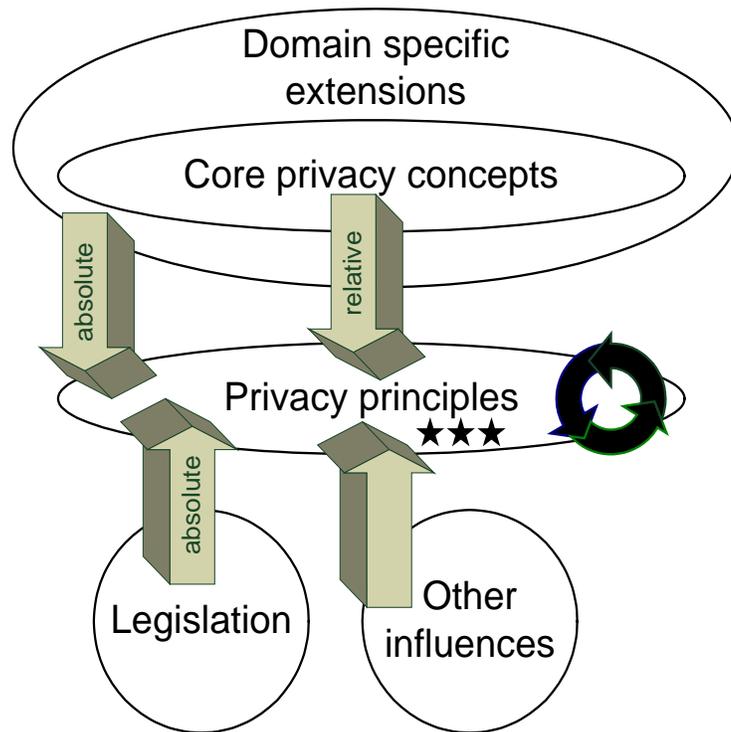
In addition to the influences assigned by the generic privacy ontology and its extensions, the privacy principles are also influenced by external factors that may be application domain independent. This is shown in Figure 4.



**Figure 4: External influences**

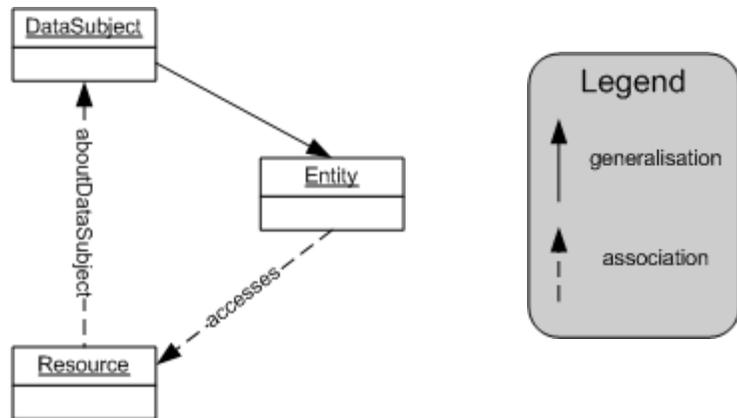
One of the examples for such an external influence is the derivation from legal documents. We cannot use legal documents in their natural language form due to the same principle problems we elaborated in Chapter 3. Basically, it is difficult to extract knowledge from natural language documents and the concepts and relationships may not be exposed easily. Furthermore, due to the number of legal and regulatory documents in the area of privacy (in any area in general), conflicts occur that cannot be solved automatically with ease and without loss of knowledge. For example, where a federal legal document may provide protection for a certain area, a state document may do the inverse and therefore a conflict exists. Unfortunately, these types of conflicts are frequent and can involve multiple simultaneously. In general, we assume that the most restrictive of the documents apply, and we do not consider the others as it is our belief that privacy protection cannot be strong enough. In order to automate the whole process, a legal ontology would need to be built that can encompass the various concepts and knowledge of legal documents, but this may never be completely achievable due to the way those documents are written - in a fairly open, imprecise (as in non-concrete) and interpretable way.

The following figure describes our conceptual approach as just discussed on a high and abstract level.



**Figure 5: Ontology big picture**

As discussed in the definition of privacy in Chapter 3, privacy is in essence about the power to control the release of personal information in line with the interests and values of the data subject. Therefore, the most basic principles are "Entity", "Resource" and "Data Subject" as defined in Chapter 3 and depicted in Figure 6. Entity is therefore a very general concept, which refers to any kind of agent, person, company or other individual. Hence, a Data Subject is a more specialized version of Entity as it has all the common properties, such as a jurisdiction by which the entity is bound. In simpler terms, a resource is defined as information about a data subject - although its formal definition is more complex as elaborated in the previous chapter.

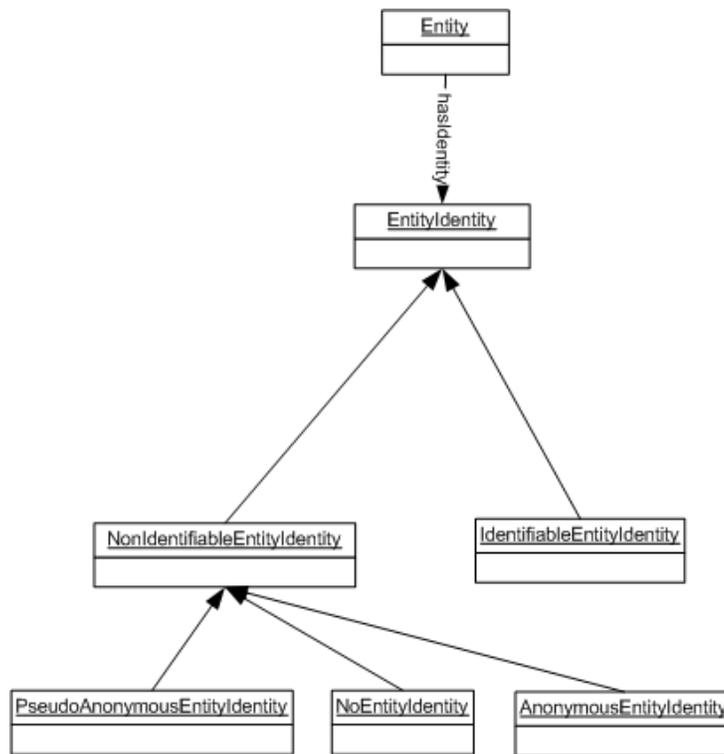


**Figure 6: Entity, Data Subject and Resource**

The idea depicted in Figure 6 is very basic and only makes the statement that a resource is about a data subject, which basically means that the resource contains personal information (to some extent) regarding the data subject. The second statement is that another entity can access that resource. This does not provide any constraints or any other requirements under which the resource may be accessed.

A data subject might not necessarily want to be known by its resource and on the other hand, the accessing entity might not want to reveal its identity to access information, perhaps of general nature only. Thus, we need to introduce the concept of identities. Both sides - the data subject and the entity accessing the resources - will want to control how they are identified - remembering that we are concerned only with business and not government transactions. The logical action is to associate a set of identities with every entity - which in turn applies to the data subject as it is an entity after all. Furthermore, an entity with multiple identities can act differently depending on the one it is using and the context in which it is used. The entity can even choose to remain anonymous if necessary. We can categorize identities themselves as either identifying the entity behind it or not. We call these two categories "IdentifiableEntityIdentity" and

"NonIdentifiableEntityIdentity" respectively. We can distinguish the latter one even further by categorizing it into a pseudo-anonymous identity and a completely anonymous one. We call them "PseudoAnonymousEntityIdentity" and "AnonymousEntityIdentity". The major difference between the latter two is that the pseudo-anonymous one can be reused multiple times, while the other cannot as it has no elements that could associate one instance with another. The anonymous one is usually used once and thus leaves no traces between different (trans)actions. We show these concepts and relationships in Figure 7.



**Figure 7: Different categories of identities**

The different categories of identities also relate to the data subject's different type or category of resources. A resource can potentially identify a data subject directly, or, even better, can identify one of the data subject's identities, which might be congruent with the data subject him- or herself. On the other hand, a resource might not identify the data subject



between the concepts and the privacy principles in Chapter 5. This is then followed by application domain specific commitments of the generic privacy ontology in Chapters 6 and 7 for two different application domains.

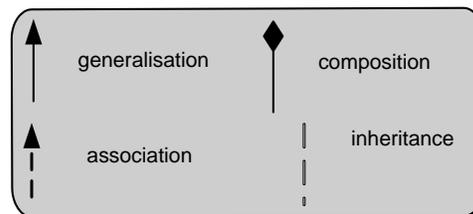
#### 4.3. Conclusion

This chapter provided a broad overview of the proposed solution and our approach to it. We discussed the idea of a generic privacy ontology and an application domain specific privacy ontology, which is reflected as an ontology commitment to the generic one. We then provided the main core concepts that are application domain independent and therefore belong to the generic privacy ontology. We also discussed the issues of external influences on the privacy principles, and an example from the conceptualisation of legal documents was described.

## 5. Generic Privacy Ontology - in details

### 5.1. Introduction

In the previous chapter, we elaborated the main idea of a generic privacy ontology and its ontology commitment for application domains. We described some of the main concepts within the generic privacy ontology, such as Data Subject, Resource and Entity. The following sections will provide a greater level of detail for all of these concepts as well as all additional concepts in the generic ontology not described in Chapter 4. Throughout the thesis, we will use the following notion:

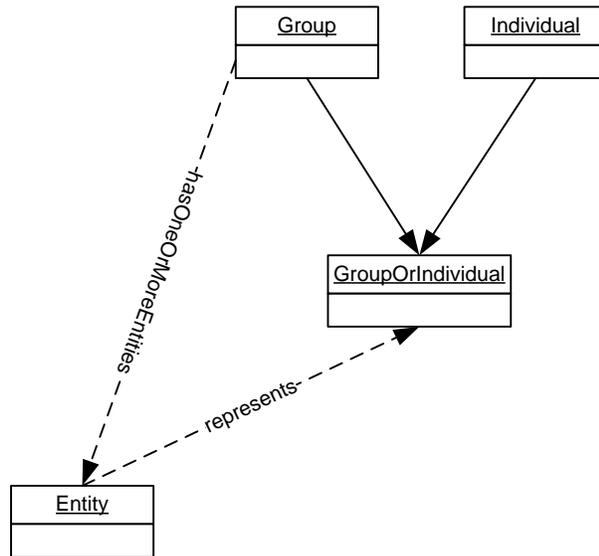


**Figure 9: Legend**

### 5.2. Ontology concepts

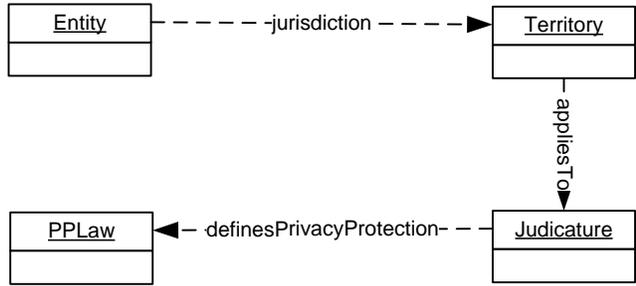
The concept of Entity was briefly described earlier. An entity represents the concept of a general and abstract agent, person, individual or company. It can also refer to a group of entities. The class entity has an association with the concept "GroupOrIndividual" as depicted in Figure 10. This association allows us to refer to a single instance of Entity during our evaluation at a later stage that is actually associated with more than one entity. This is necessary if an entity consists of multiple entities. We have to use an association here as we further refine (specialize) the concept of entity further at a later stage and both Group and Individual are already specializations of GroupOrIndividual. Hence, the group concept is just a

container that can hold multiple entities, which can be cascaded if necessary. The concept of Individual does not only refer to human beings, but can include any type of real world or digital agent, person or provider.



**Figure 10: Entity and GroupOrIndividual concepts**

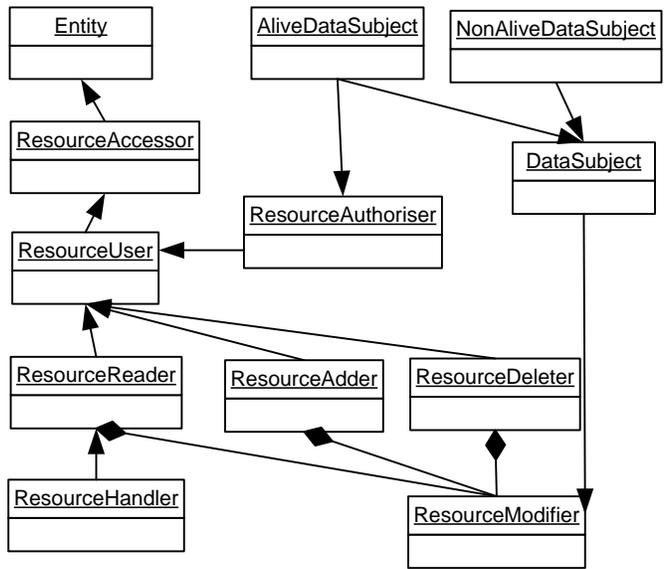
Every entity has an association named "jurisdiction" that defines to which territory an entity belongs - see Figure 11. Territory therefore refers to a specific country or region of the world in most of the cases. An entity can belong to one or more territories at any given time. However, if, for example, an entity is a multi-national one that operates internationally and multiple distinct territories would apply, we simply use the grouping mechanism from Figure 10 and group them together such that a separate instance appears for every occurrence of that entity in every territory. Every territory has specific judicatures, and hence its own particular privacy protection laws and regulations that apply. These significantly influence the level of privacy protection. As it is beyond the scope of this thesis to develop an ontology for privacy protection laws, we assume that it can be done in general but, in the following chapters, we refer only to a few examples that have been developed for this particular purpose.



**Figure 11: Entity and Territory**

### 5.2.1. Entity hierarchy

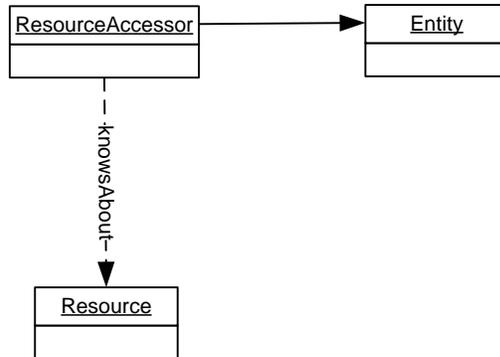
In most cases, an entity is accessing or processing personal information of a data subject. Therefore, we have to define various subclasses of entities to deal with different access types or privileges and the implications for the privacy principles with regards to that type of access. The whole hierarchy of is shown in Figure 12.



**Figure 12: Entity hierarchy**

The first level down from the entity is classified as "ResourceAccessor", which represents the concept of an entity that will deal with personal information at some stage. Hence, the accessor cannot read, alter or delete

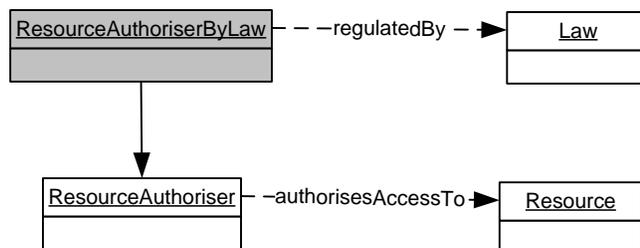
and information, but can only pass it on to others. The accessor has knowledge about having personal information, but knows neither the content nor to whom it belongs as the content is unavailable - see Figure 13. An example of such an accessor in real world terms would be the mailman in the real world or an email service provider in the digital world.



**Figure 13: ResourceAccessor**

A "ResourceUser" is the next logical step down in the hierarchy. This concept classifies an entity as knowing the identity of the data subject. The identity may or may not be congruent with the actual data subject. In most cases, a resource user is also a recipient of personal information, but cannot access that personal information as such.

A "ResourceAuthoriser" is an entity that can control the release of personal information on behalf of a data subject. This can be seen as power of attorney for multiple reasons.



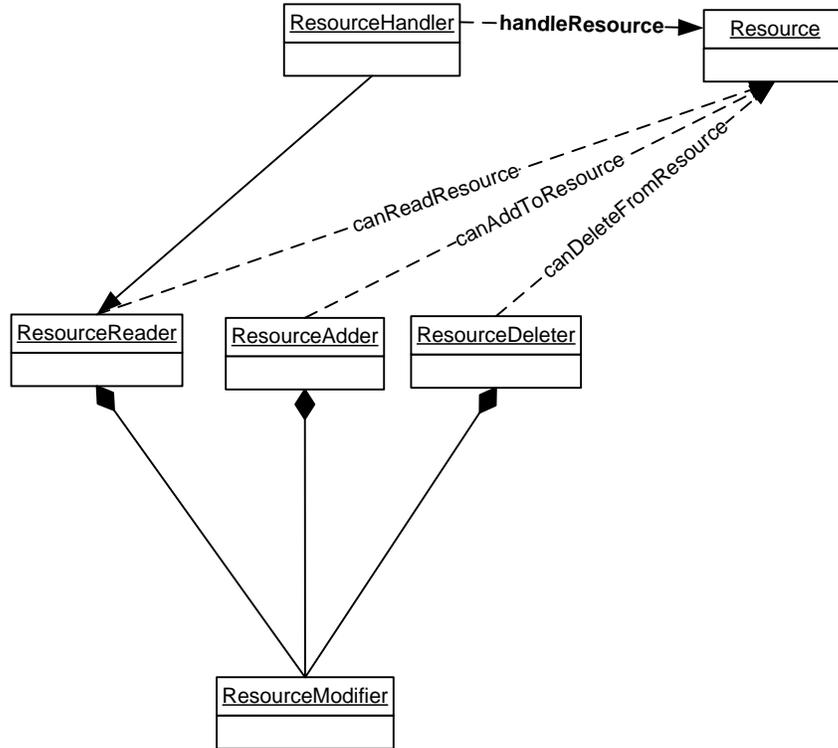
**Figure 14: ResourceAuthoriser**

Reasons include legal requirements as well as power in the case where the data subject is unfit or unable to control it him- or herself. For example, this could be a parent controlling the release of personal information about a minor / child. The authoriser does not necessarily have the ability to read or alter the personal information controlled. However, as it needs to know about the identity of the personal information, it is only logical that it be a specialization of the ResourceUser. The actual process of sharing personal information is described below. Furthermore, a specialized type of the ResourceAuthoriser exists as depicted in Figure 14, which is called "ResourceAuthoriserByLaw". This particular concept is used in cases where legal requirements demand the control of personal information. The legal requirements are modeled by the different laws and regulations, but we simply refer to the actual law without proper semantic modeling as this would exceed the scope of this thesis.

The concept of "DataSubject" is one of the core concepts in this ontology. Resources are about data subjects in general and it is the desire of the data subjects to control their own personal information - the resources - consistent with own their values as described in the definition of privacy in Chapter 3. As different laws in different countries describe privacy protection differently, it is necessary to distinguish between data subjects that are "alive" and data subjects that are not alive. Therefore, "alive" is used in the context of human beings only as privacy protection by law varies according to the living state of a person. For example, in some countries, privacy protection laws cease to apply with the death of a person. In the digital world, we use the term similarly as, in general, digital data subjects represent human beings at some point. If the entity is not a human being but a corporation, we assume it to be alive as long as it exists in its form. The second implication of an "alive" data subject is the ability to

control personal information, which can no longer be done by someone who is dead. Therefore, an "alive" data subject is also a ResourceAuthoriser. This is ultimately true in most cases, as it is the data subject who is to control his or her own information. A number of cases exist where this is not true, for example if the data subject is legally or otherwise unfit to do so as explained in the previous section.

So far, it is not actually possible to access personal information and be able to read or alter the content, apart from the data subject, who is able to do so in most cases. The concept of "ResourceReader" represents an entity that knows about the identity of a data subject (which can be pseudo-anonymous or fully anonymous) and can access some or all parts for certain purposes and under certain conditions. As these conditions and purposes are merely resources or meta-resources, we describe them below in greater detail. In general however, the conditions and purposes of reading personal information are always attached to it directly. A ResourceReader does not have the ability to alter personal information; this is done by the entity represented by the concepts of "ResourceAdder" and "ResourceDeleter" respectively. While the former can only amend information, the latter can only delete it. Combining all of the concepts of "ResourceReader", "ResourceAdder" and "ResourceDeleter" leads to the concept of "ResourceModifier", which is an entity able to modify resources under certain conditions and for certain purposes as defined by a "ResourceAuthoriser" or the data subject as depicted in Figure 15.



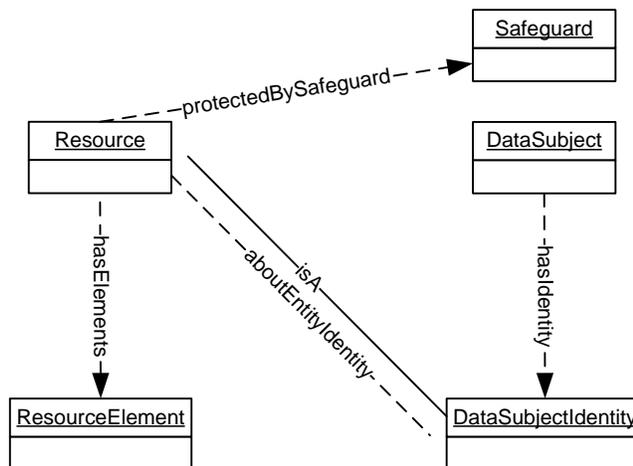
**Figure 15: ResourceHandler and Modifier**

We have an additional concept in this hierarchy, which is called "ResourceHandler". Such an entity is a specialized type of a "ResourceReader". The entity represented by this concept has the ability to read personal information under certain conditions and for certain purposes, as described above, but can also transform them without altering the actual content. This is useful in cases where personal information need to be translated from one format to another, such as from one language to a different one, or from a physical form into digital one or vice versa. The ResourceHandler has no other privileges and cannot reveal personal information to any other entity. This may seem to be contradictory as an interpreter, for example, translates information and gives it to someone else. However, we regard this as the interpreter translating it for the data subject and the data subject then forwarding it to other entities. A ResourceHandler therefore has additional attributes,

which are trust and reputation as well as quality. The former ones refer to the trust and belief in reputation that a data subject has for the handler with regards to handling his or her information, while the latter refers to the actual quality the handler delivers and the accuracy of the outcome. This is important in particular where transformations cannot be done with one hundred percent accuracy and it is up to the handler to perform well.

### 5.2.2. Resources

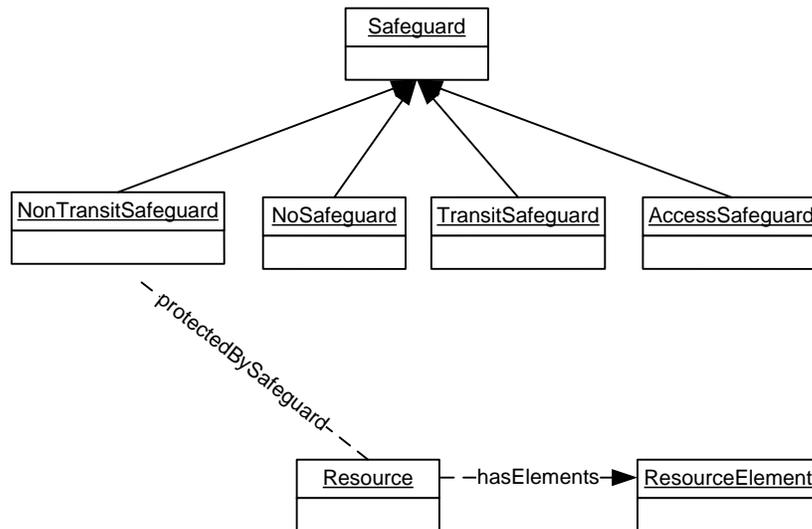
We described the resource concept in Chapter 4, but will provide more detail here. The concept of resource is an abstract one that identifies the abstract concept of DataSubjectIdentity as depicted in Figure 16.



**Figure 16: Resource, Identity and Safeguard**

A resource is associated with a number of ResourceElements that contain the actual information. They, in turn, can be used to identify certain identities as described earlier. Every resource is associated with and hence protected by the concept of "Safeguard". The ontology caters for four types of safeguard categories: "TransitSafeguard", "NonTransitSafeguard", "AccessSafeguard" and "NoSafeguard". The concept of resource is associated with a "NonTransitSafeguard" as shown in Figure 17, as a

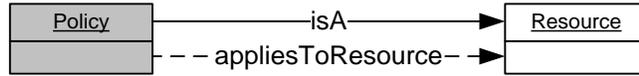
resource refers to data that is not in transit. Examples for such a "NonTransitSafeguard" may be "AES-Encryption" in the digital world or "A4-Envelope" in the real world. Both are concepts of storage safeguards that prevent unauthorized access to the information they protect at various levels of protection. The level of protection that a particular safeguard provides is reflected by the security privacy principle and is part of the evaluation process determining the level of privacy preservation. In the example above, we would consider "AES-Encryption" as a safeguard with a protection level of "high", while an envelope may be awarded a protection level of "low". As with all other concepts, this mechanism is extendible and new safeguards can be added as needed and as technology evolves. We will provide further examples of safeguards throughout the chapter as deemed appropriate.



**Figure 17: Resource and Safeguard**

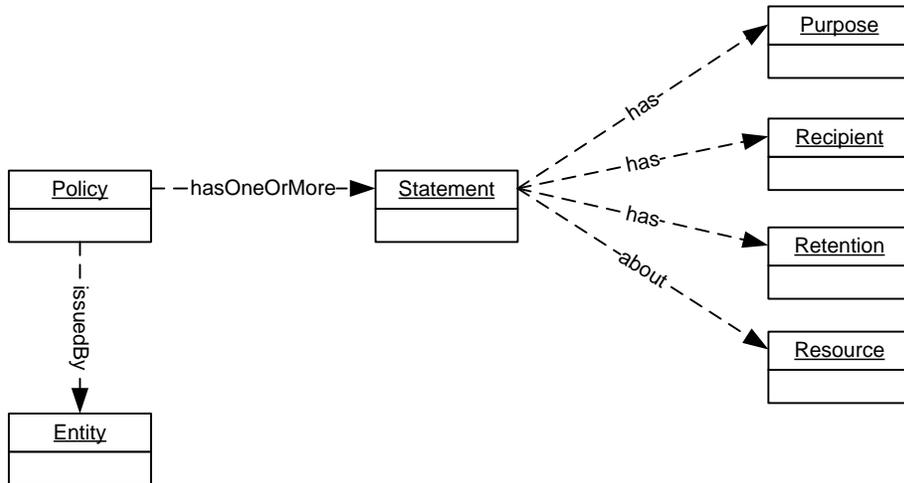
However, security is only one part of privacy protection. A second part is the purpose for which personal information has been collected. This is modeled by the concept of policy. A policy is a specialized type of resource and therefore is usually protected by safeguards as well. In addition and from an abstract point of view, a policy applies to the concept of resource

as shown in Figure 18. Although a policy is a resource, it does not usually apply to itself, meaning the policy and the resource with which it is associated with are disjoint.



**Figure 18: Policy and Resource**

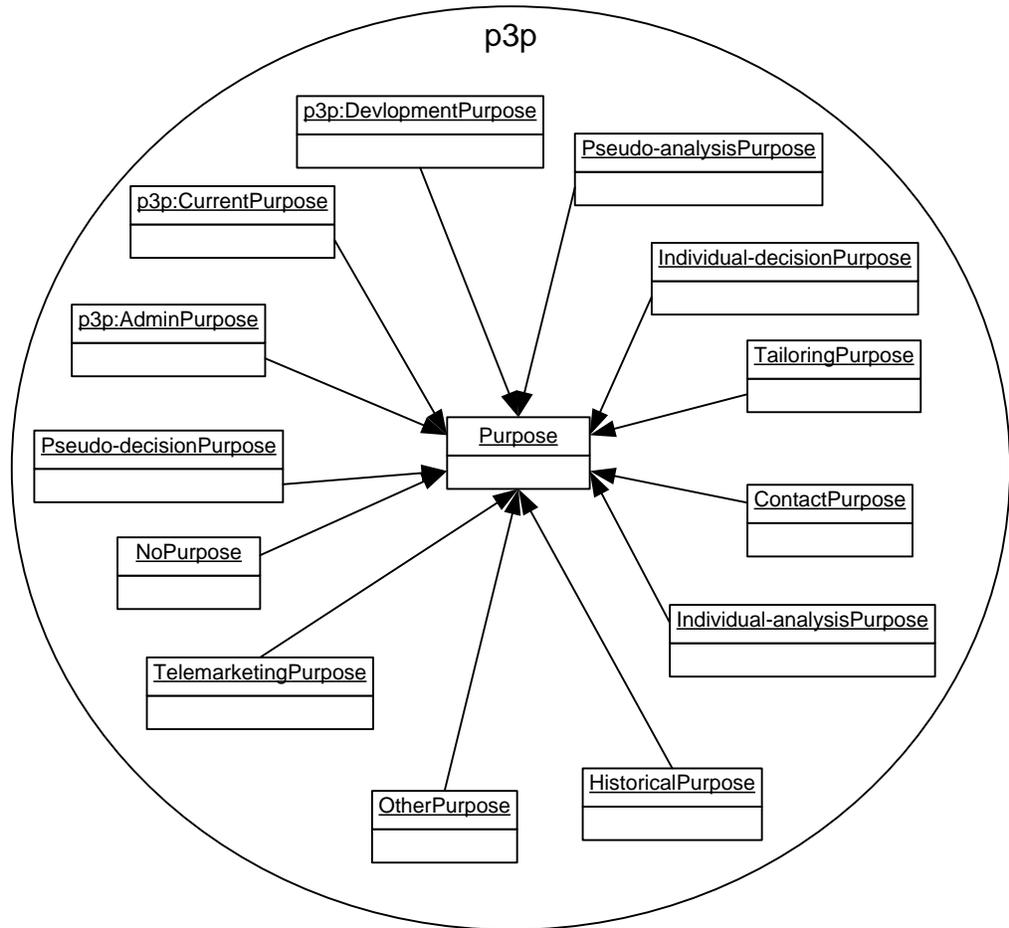
A policy in the ontology is more complex, however. A policy is defined and issued by a particular entity and has a set of statements that on their own have one or more purposes, one or more recipients, a certain retention and are about a resource (see Figure 19). Examples for purposes are numerous and we have used P3P as an example to determine some categories and show them in Figure 20.



**Figure 19: Policy**

We define the concept of Repository as a container that contains a number of resources and one or more associated policies. This is useful to store the policies under which certain personal information has been collected

together with the actual information. This helps to ensure that intended use is congruent with the purpose.

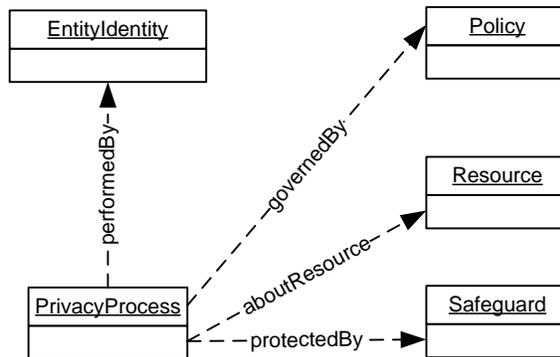


**Figure 20: Purposes in the P3P domain**

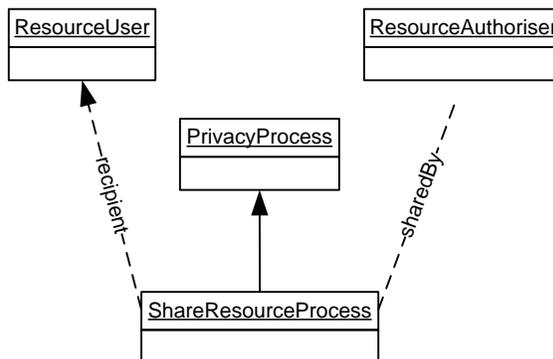
### 5.3. Privacy Processes

The generic privacy ontology contains the concept of privacy processes. These processes represent different tasks or transactions that occur frequently when processing personal information and can be extended when necessary. A PrivacyProcess is defined as an abstract concept that is about a resource and performed by a particular identity of an entity, governed by a particular policy and protected by a safeguard as shown in Figure 21. Actual processes derive from this basic conceptualization and

add their own associations. An example of such a process is the "ShareResourceProcess" as shown in Figure 22. In addition to the elements above, it has an association with the ResourceAuthoriser (e.g. the data subject) who is permitted to share the resources, and an entity as the recipient of the data. As an entity can be an individual or a group, simultaneous sharing with multiple parties is possible. However, the conditions of sharing may require the involvement of multiple sharing processes if entities reside in different territories or have other properties that may differ from each other. This can then be governed by different policies in the privacy process.



**Figure 21: Privacy Process**



**Figure 22: ShareResource Process**

#### 5.4. Detailed illustration

Before we describe the privacy principles and the influence of the various concepts in the ontology, we will describe a few concepts in greater detail. So far, we have omitted most of the attributes of most of the concepts in order to limit ourselves to the ideas behind them and avoid too many technical details. However, we will show some technical details now for some concepts, while the complete set of technical details can be found within our implementation of the ontology, written in OWL. Some more pseudo code snippets will also be provided in the following chapters to assist with the understanding. However, the basic principles and ideas should be clear without these technical details, which may distract from, rather than assist the discussion. The examples below are described in an implementation-independent manner as we describe our choice of implementation language in Chapter 8.

As a first example, we highlight the "Entity" concept:

The concept of entity has a single attribute named "id", which is an arbitrary identifier that identifies instances of this concept. We choose an arbitrary value here as an entity is not supposed to have any identifying characteristics of the actual individual. Entity also has three associations. The first one is named "represents" with a cardinality of one and associates the concept of "GroupOrIndividual" with this concept. The second association is named "isInJurisdiction" with a cardinality of one or more and associates Entity with the concept of "Territory", which defines the governing country or region(s) by which this entity is bound. As the cardinality allows for more than one, it supports the entity being located in multiple regions at the same time, for example a particular state and a particular country (in which that state is located). This then allows us to apply different privacy regulations from the various territories. The third

association is named "hasIdentity" and is an association with a cardinality of one or more and associates Entity with the concept of "Identity", meaning that every identity can have an arbitrary number of identities.

Our second example describes the concept of Identity:

The main concept of Identity is abstract and contains an arbitrary identifier "id" only. A specialization or subclass of Identity is "IdentifiableEntityIdentity", which contains several attributes that uniquely identify the entity behind them. Its first attribute "id" is inherited from its super-class "Identity" and used in the same way as before. The second attribute is named "name" and is a string with a cardinality of one that describes the actual and real name of this identity and hence, entity. The third attribute is named "dateOfBirth" and has a cardinality of zero or one and describes the date of birth. We use a cardinality of zero or one here to allow for application domain commitments to use identity for both persons (having a date of birth) and non-person (e.g. companies, not having a date of birth). The fourth attribute is "locality" and has a cardinality of one and describes the location of the identity. It is noteworthy that this is not the same as the territorial association of the entity concept.

In general however, most of the concepts in the generic privacy ontology have as few attributes as possible in an attempt to make it as application domain-independent as possible and the commitments add attributes as required. The only attribute that is available to all concepts is "id", which, as shown, is an arbitrary identifier for the particular instance of that concept.

## 5.5. The Privacy Principles

We have listed the privacy principles in previous chapters, but provide a more detailed description here. This includes the various quality assessment aspects that are used to determine the level of privacy experienced. The top level of our evaluation contains the quality aspect that is evaluated. On the second level, this is followed by the different quality assessment criteria, which in turn have certain possible assignments - shown as third level items. The assignments themselves do not qualify the outcome of any privacy preservation evaluation process as such. The individual influence measurement determines this.

One of the privacy principles, "Intention and Notification", is distinct from this notion as it is usually not used during every transaction that involves personal information, but is more a preliminary requirement in certain countries to start collecting personal information. Therefore, although we have listed it below, it is not actively used by the ontology as such. However, data collectors and processors may need to adhere to the different aspects of this principle individually as deemed appropriate.

#### 5.5.1. Data Quality (Quality Aspect 1)

##### 5.5.1.1. Adequate (Quality assessment criteria 1)

5.5.1.1.1. Yes, fully: ★★★★★

5.5.1.1.2. No / Not stated: ★

##### 5.5.1.2. Relevance to purpose (Quality assessment criteria 2)

5.5.1.2.1. Very High: ★★★★★

5.5.1.2.2. High: ★★★★

5.5.1.2.3. Medium: ★★★

5.5.1.2.4. Low: ★★

5.5.1.2.5. None / Unknown: ★

##### 5.5.1.3. Correctness (Quality assessment criteria 3)

- 5.5.1.3.1. Very High: ★★★★★
- 5.5.1.3.2. High: ★★★★
- 5.5.1.3.3. Medium: ★★★
- 5.5.1.3.4. Low: ★★
- 5.5.1.3.5. Not at all / Unknown: ★

5.5.2. Security (Quality Aspect 2)

5.5.2.1. Safeguards adequate (Quality assessment criteria 1)

- 5.5.2.1.1. Very high: ★★★★★
- 5.5.2.1.2. High: ★★★★
- 5.5.2.1.3. Medium: ★★★
- 5.5.2.1.4. Low: ★★
- 5.5.2.1.5. None / Not stated: ★

5.5.2.2. Security policy adequate

- 5.5.2.2.1. Yes, fully - ★★★★★
- 5.5.2.2.2. No / Not stated ★

5.5.2.3. Data destruction policy adequate

- 5.5.2.3.1. Yes, fully - ★★★★★
- 5.5.2.3.2. No / Not stated ★

5.5.2.4. Contingency plan adequate

- 5.5.2.4.1. Yes, fully - ★★★★★
- 5.5.2.4.2. No / Not stated ★

5.5.2.5. Personnel requirements adequate

- 5.5.2.5.1. Yes, fully - ★★★★★
- 5.5.2.5.2. No / Not stated ★

5.5.2.6. Privacy enhancing technologies adequate

- 5.5.2.6.1. Yes, fully - ★★★★★
- 5.5.2.6.2. None used / Not stated ★

5.5.2.7. ICT infrastructure adequate

- 5.5.2.7.1. Yes, fully - ★★★★★

5.5.2.7.2. No / Not stated ★

5.5.3. Data subject's rights

5.5.3.1. Access privileges to own data

5.5.3.1.1. Full: ★★★★★

5.5.3.1.2. Most: ★★★★★

5.5.3.1.3. Some: ★★★

5.5.3.1.4. Few: ★★

5.5.3.1.5. None / Unknown: ★

5.5.3.2. Level of ability to request

rectifications/supplementations/deletions

5.5.3.2.1. Very High: ★★★★★

5.5.3.2.2. High: ★★★★★

5.5.3.2.3. Medium: ★★★

5.5.3.2.4. Low: ★★

5.5.3.2.5. None / Not Stated: ★

5.5.3.3. Ability to block content for certain purposes

5.5.3.3.1. Yes: ★★★★★

5.5.3.3.2. No / Not stated: ★

5.5.3.4. Ability to object against processing

5.5.3.4.1. Yes: ★★★★★

5.5.3.4.2. Partly: ★★★

5.5.3.4.3. No / Not stated: ★

5.5.4. Legitimate Grounds of Processing

5.5.4.1. Unambiguous consent

5.5.4.1.1. Yes ★★★★★

5.5.4.1.2. No ★

5.5.4.2. Processing to fulfill contract requirement

5.5.4.2.1. Yes ★★★★★

5.5.4.2.2. No ★

5.5.4.3. Legal reasons

5.5.4.3.1. Yes - refer to reason ★★★★★

5.5.4.3.2. No ★

5.5.4.4. Protection of vital interest of data subject

5.5.4.4.1. Yes - refer to reason ★★★★★

5.5.4.4.2. No ★

5.5.4.5. Data belongs to following sensitive category

5.5.4.5.1. Religion: ★

5.5.4.5.2. Philosophical beliefs: ★

5.5.4.5.3. Race: ★

5.5.4.5.4. Political Opinions: ★

5.5.4.5.5. Medical Data: ★

5.5.4.5.6. Sex life: ★

5.5.4.5.7. Trade union memberships: ★

5.5.4.5.8. Data about criminal convictions or unlawful  
behaviour: ★

5.5.4.5.9. No: ★★★★★

5.5.5. Transparency

5.5.5.1. Data obtained directly notify of processing

5.5.5.1.1. Before collection: ★★★★★

5.5.5.1.2. During collection: ★★★

5.5.5.1.3. After collection: ★★

5.5.5.1.4. Never: ★

5.5.5.2. Data obtained indirectly notify of processing

5.5.5.2.1. Source of data known: ★★★★★

5.5.5.2.2. Source of data unknown: ★

5.5.5.3. Identity of processor revealed

5.5.5.3.1. Yes: ★★★★★

5.5.5.3.2. No: ★

- 5.5.5.4. Purpose of data stated
  - 5.5.5.4.1. yes: ★★★★★
  - 5.5.5.4.2. partly: ★★★
  - 5.5.5.4.3. no: ★
- 5.5.5.5. Recording in accordance with law
  - 5.5.5.5.1. Yes - refer to law(s): ★★★★★
  - 5.5.5.5.2. No - specify reason: ★
- 5.5.5.6. Third party involved
  - 5.5.5.6.1. yes - specify which ★★★
  - 5.5.5.6.2. no: ★★★★★
- 5.5.5.7. Ability to object to involve third party
  - 5.5.5.7.1. Yes: ★★★★★
  - 5.5.5.7.2. No: ★
- 5.5.6. Finality principle
  - 5.5.6.1. Level of purpose specified
    - 5.5.6.1.1. Very High: ★★★★★
    - 5.5.6.1.2. High: ★★★★★
    - 5.5.6.1.3. Medium: ★★★
    - 5.5.6.1.4. Low: ★★
    - 5.5.6.1.5. None / Not Stated: ★
  - 5.5.6.2. Purpose legitimate
    - 5.5.6.2.1. Yes: ★★★★★
    - 5.5.6.2.2. No: ★
  - 5.5.6.3. Retention period
    - 5.5.6.3.1. None: ★★★★★
    - 5.5.6.3.2. As long as needed to fulfill contract: ★★★★★
    - 5.5.6.3.3. Specified period - anonymized: ★★★★★
    - 5.5.6.3.4. Specified period - not anonymized: ★★★
    - 5.5.6.3.5. Forever - anonymized: ★★

- 5.5.6.3.6. Forever - not anonymized: ★
- 5.5.7. Processing by a third party - data sharing
  - 5.5.7.1. Instructed by controller
    - 5.5.7.1.1. Yes: ★★★★★
    - 5.5.7.1.2. No: ★
  - 5.5.7.2. Level of compliance with obligations of controller
    - 5.5.7.2.1. Very High: ★★★★★
    - 5.5.7.2.2. High: ★★★★
    - 5.5.7.2.3. Medium: ★★★
    - 5.5.7.2.4. Low: ★★
    - 5.5.7.2.5. None / Not Stated: ★
  - 5.5.7.3. Legal binding contract in place
    - 5.5.7.3.1. Yes: ★★★★★
    - 5.5.7.3.2. Partly: ★★★
    - 5.5.7.3.3. No / Unknown: ★
- 5.5.8. Accountability
  - 5.5.8.1. Person nominated to watch over compliance
    - 5.5.8.1.1. Yes: ★★★★★
    - 5.5.8.1.2. No / not stated: ★
- 5.5.9. Openness
  - 5.5.9.1. Policies about procedures available
    - 5.5.9.1.1. Yes - refer to location: ★★★★★
    - 5.5.9.1.2. No / Not stated: ★
- 5.5.10. Anonymity
  - 5.5.10.1. Data anonymized
    - 5.5.10.1.1. Yes, fully: ★★★★★
    - 5.5.10.1.2. Yes, pseudo-anonymized: ★★★
    - 5.5.10.1.3. No / Unknown: ★
- 5.5.11. Consent

5.5.11.1. Explicit

5.5.11.1.1. Standing: ★★★★★

5.5.11.1.2. Other: ★★★

5.5.11.1.3. No: ★

5.5.11.2. Implicit

5.5.11.2.1. National Security: ★★★

5.5.11.2.2. Legal Obligation: ★★★

5.5.11.2.3. Protection of Vital Interests of Data Subject:

★★★★★

5.5.12. Transfer between different jurisdictions

5.5.12.1. If transfer - **justification** to transfer data to jurisdiction with different privacy protection laws

5.5.12.1.1. unambiguous consent by data subject - refer to consent: ★★★★★

5.5.12.1.2. necessary for fulfillment of contract - refer to contract: ★★★★★

5.5.12.1.3. legal requirement - refer to law: ★★★★★

5.5.12.1.4. None: ★

5.5.12.1.5. Not stated: ★

5.5.12.2. Privacy Protection Laws **Standards**

5.5.12.2.1. Very High: ★★★★★

5.5.12.2.2. High: ★★★★★

5.5.12.2.3. Medium: ★★★

5.5.12.2.4. Low: ★★

5.5.12.2.5. None / Not Stated: ★

## 5.6. Relationship between concepts and principles

So far, we have discussed different concepts and relationships in the generic privacy ontology as well as the privacy principles, which are also called quality aspects in the context of privacy for our purposes. In this section, we explain how the various privacy concepts of the generic privacy ontology map to these privacy principles. Every concept in the generic privacy ontology has an association with one of the quality aspects (i.e. the privacy principles) to state that it influences it in some way. We understand that this is a very fuzzy terminology, but the actual influence will be decided by the privacy ontology commitment as determined by the application domain expert. For example, every resource and its dependent purpose influences the principle or quality aspect of data quality. However, as the actual type of resource or purpose is not determined by the generic privacy ontology, we can only say that it influences that quality aspect (data quality) in some way. Therefore, data quality is a function of resource and purpose. When it comes to ontology commitment, a domain expert would specify multiple types of resources and purposes and therefore have to specify the data quality aspect as well. The domain expert has to follow the structure of that particular quality aspect and determine appropriate values for the various quality assessment criteria. In our example, we assume that our domain expert creates a privacy ontology commitment for the medical domain and classifies "patient information" as a particular resource within the resource tree and purpose "ProvideMedicalAttention" within the purpose hierarchy. One of the quality assessment criteria of data quality is "Data Adequacy". The domain expert would classify this with 5 stars from our range of possible ordinal values between 1 and 5, where 5 is the highest. We therefore conclude that the data provided here is adequate to a high level, which contributes to our evaluation of the level of privacy received in section 5.7.

A second example to describe our approach is the very generic concept of "Entity". Entity itself does not have any influence by itself, although, one of its associations does. Entity has an association named "isInJurisdiction" with the concept of "Territory". Territory in turn has an association with privacy principle or quality aspect 5.5.12 (transfer between different jurisdictions). The domain expert would then have to determine the specific commitment. In our example, we assume the application domain is e-commerce and the domain expert needs to model certain business processes. With regards to the territory and its mapping to the quality aspect, the domain expert would have to assign values to the various territories in which customers and clients may reside, in context of the application domain. For example, if the instance "Europe" were one of the territories, a value of "5 stars" might be assigned and a value of "3 stars" to "United States of America". During the privacy evaluation process, the level of privacy would be determined according to the values assigned and remain the same, for example, if both entities were in the same territory. They would drop if the data subject were in a 5-star territory while the other entity were in a 3-star one, but would remain the same if the data subject were in a 3-star one and the other entity in a 5-star one. This is true as the communication channels and laws of both territories apply and therefore we have to use the minimum of both.

#### 5.7. Privacy evaluation process

So far, we have discussed the various concepts, relationships, privacy principles and how they are associated with each other. This section describes the privacy evaluation process, which is the process that determines the level of privacy experienced when dealing with or performing certain transactions or dealing with certain systems.

In the previous section, we described how generic ontology concepts map to various privacy principles and how a domain expert for a particular application domain would assign ordinal values. We have decided to use ordinal values as privacy is an endogenous conception and not mathematically precise. In addition to the various levels of privacy as determined by the domain expert at class level or during run-time, preferences can be taken into account. Every quality aspect has a default weight assigned to it, depending on the application domain. This can be used by the user to tailor the level of privacy experienced to his/her own needs. For example, a user might not care that his or her personal information are taken offshore to a different country with weaker privacy protection laws. The weight of this quality aspect would then be reduced accordingly to reduce the impact factor of the overall evaluation. Hence, we use the following formula to define the level of privacy  $L_p$ :

$$L_p = Round \left( \frac{\sum_{i=1}^n w_i \times Q_i}{\sum_{i=1}^n w_i \times max} \times max \right)$$

where  $n$  is the number of quality aspects assessed,  $w$  the weight or impact of a particular privacy principle and  $Q_i$  the privacy evaluation of the particular principle. The weight  $w$  is defined as:  $0 \leq w \leq 1$ , where 0 is defined as having no impact or no importance on that particular quality aspect and 1 having full impact. Max is an ordinal value, which we have defined as number 5, which is equivalent to the highest value a quality aspect can be assessed at. The rounding function Round rounds the value up or down to the next ordinal value. We assign the same number of stars to the outcome evaluation that this number dictates. The formula for  $Q_i$  is defined as:

$$Q_i = \frac{\sum_{j=1}^m A_j}{m}$$

where  $m$  is the number of quality assessment criteria for  $Q_i$  and  $A$  the value that has been determined by the application domain expert for a particular assessment criteria and is an ordinal value that ranges between 1 and max, which is defined as number 5 in this thesis. It is important to say that we expect the weight values to be fairly static across any given application domain as this is determined by a domain expert once and should apply to a large number of individuals within that domain. Nevertheless, individuals may change the weight for a particular domain to tailor it to their own needs. For example, we have determined the impact factor of the safeguard privacy principle as 0.5 as safeguards cannot be perfect in that domain due to the vital protection of the interests of the data subject, which means that it needs to be possible to access medical details in an emergency which is possible if safeguards are not too strong, but as they are less important in this domain, the level of privacy does not deteriorate to an inappropriately high level.

Our evaluation of the level of privacy experienced can be done on a concept by concept or process by process basis, where each is assessed individually. However, if many concepts and processes are involved, we state that the minimum of the individual assessments for each privacy principle applies. For example, if we have evaluated privacy principle "Safeguards" three times within a certain system or transaction and the outcome varies each time (which is likely), then we would use the lowest of those assessments to determine the overall value.

## 5.8. Conclusion

In this chapter, we described the conceptualization of the generic privacy ontology, its concepts and relationships as well as the privacy principles. Furthermore, we elaborated the level of privacy for particular concepts relating to the privacy principles and how the level of privacy is assessed when performing certain transactions.

The generic privacy ontology is composed of virtual classes that have only a few attributes which remain abstract. More concreteness can be added by a domain expert who can extend the generic privacy ontology to a particular application domain and, at that level, attributes become more concrete. This is in contrast to the instance level of a particular application domain privacy ontology, where the concepts represent particular individuals. The concrete class is therefore specified as a collection of instances.

In the next two chapters, we provide an application domain extension of this generic privacy ontology to more concrete classes and thus attributes for two different domains. This will also allow us to show the actual impact values for various concepts which in turn allow us to determine the level of privacy experienced.

## 6. Specialization I - Restricted medical domain

### 6.1. Introduction

In this chapter, we extend our discussion beyond the generic privacy ontology and describe its application to a restricted medical domain. As the medical domain is vast and has countless concepts related to a variety of issues that have no privacy implications, we have to limit ourselves to keeping the example concise and ensuring that only privacy-related concepts are included. We have selected a hospital situation and will first provide the concepts, relationships and processes for that domain. We have chosen a real world scenario as an example to show that the ontology can be applied to both real and digital world scenarios, thereby emphasizing its abstraction from any application domain as done in some of our previous work [69]. This is followed by the extension of the generic privacy ontology with concepts from the medical domain to create the medical privacy ontology. We then describe the example and how the level of privacy is influenced, computed and evaluated and then conclude this chapter.

### 6.2. Medical Domain Ontology

Since this thesis focuses on the issue of privacy, we will describe an application domain ontology from the medical domain that is a very small subset of concepts from the total medical domain that are actually related to the issue of privacy. A domain expert would proceed similarly as only those concepts that are going to be used are actually part of an ontology, in order to reduce complexity. The most commonly used ontology in the medical domain is the Unified Medical Language System (UMLS). However, for our elaboration and in order to avoid confusion, we will present our

own restricted application domain ontology for health. This is important as the majority of concepts in UMLS or the medical domain in general are not applicable to the context of privacy. For example, the medical domain has a very large number of concepts that deal with different illnesses, their treatments and causes as well as drug uses and drug composition, which are of no interest to use as they do not contribute to our elaboration. In general, we are not limited to creating ontologies from scratch, but can integrate and use other ontologies with a reasonable amount of effort by a domain expert, who would choose appropriate concepts from the ontology and create a sub-ontology of the main one [68].

We present the domain ontology in an explanatory fashion to make it easier to understand the concepts and follow the example. Furthermore, we do not claim this to be a complete or concise ontology for this application domain.

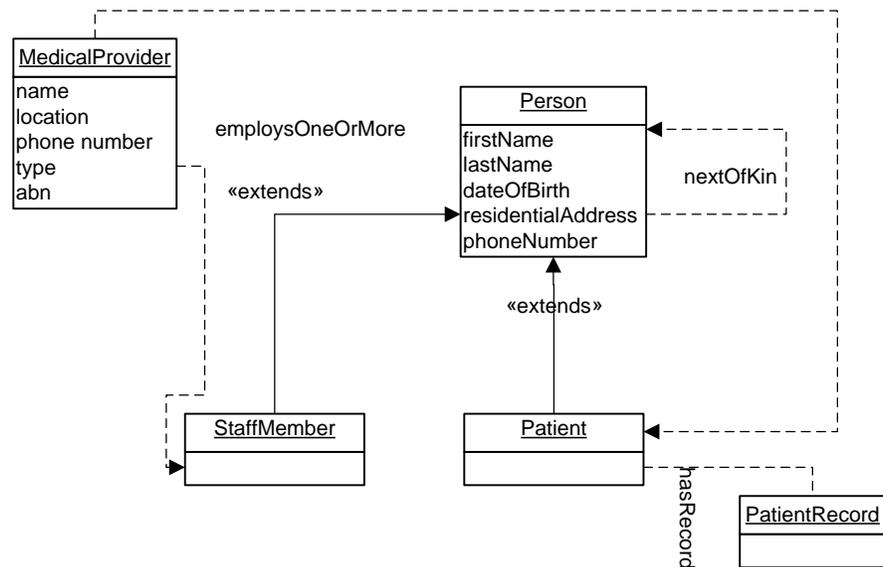
In the medical domain as in most other application domains, the concepts of resources and entities are the most important and the most common.

#### 6.2.1. Concepts

The most important concept in our clinical example is "Person", which we define as a natural person who is not deceased. A person would also have certain demographic details, namely first name, last name, date of birth, residential address, phone number. We would also have next of kin, which is an association to a different "Person". As this is not specific to the medical domain, we could reuse this concept from an ontology of the demographic domain if existent. We then specialize the concept of person with the concept of "Patient", which we define as a person, who is under medical care or treatment due to a medical issue. A different type of

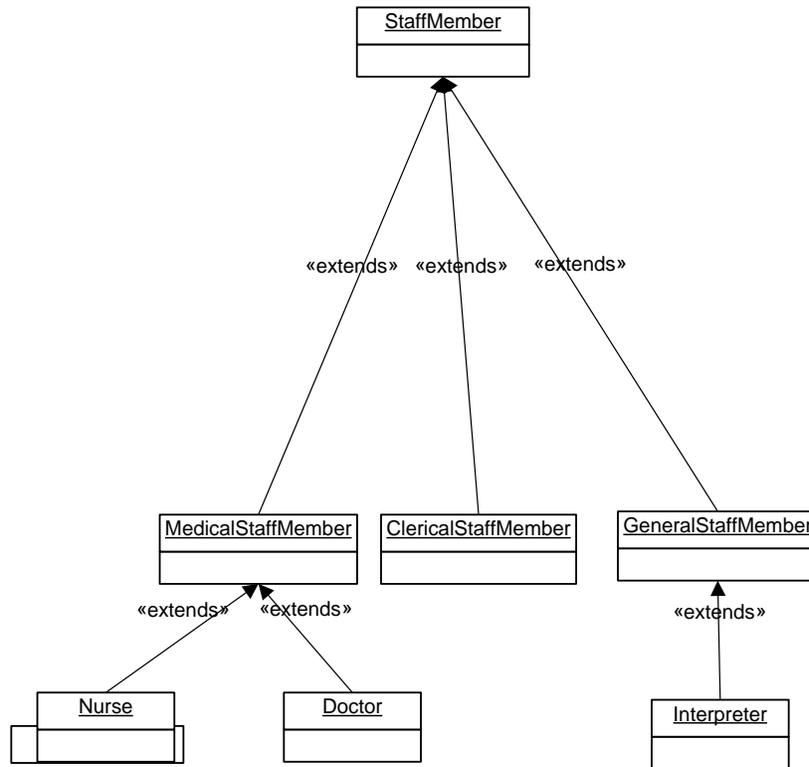
person would be the concept of "StaffMember". A staff member can be specialized much further, but we have limited ourselves to the concepts of "MedicalStaffMember", "ClericalStaffMember" and "GeneralStaffMember" for the sake of clarity here. Medical staff can be broken down into "Doctor" and "Nurse", but may have other concepts that we have also omitted here as they do not play any role in our elaboration. Another important concept is "MedicalProvider" which employs a number of staff and which we will use here as the entity that takes care of patients. Both conceptualizations are depicted in Figure 23 and Figure 24.

takesCareOfOneOrMore



**Figure 23: Medical Ontology subset 1**

A patient record is a collection of personal information with different levels of sensitivity and access privileges. It has to be categorized appropriately to allow sufficient access and privacy protection. We assume that a patient record can contain demographic information, information about chronic conditions such as allergies, previous diagnosed medical conditions, previous medical care received as well as information about medications or current treatments.

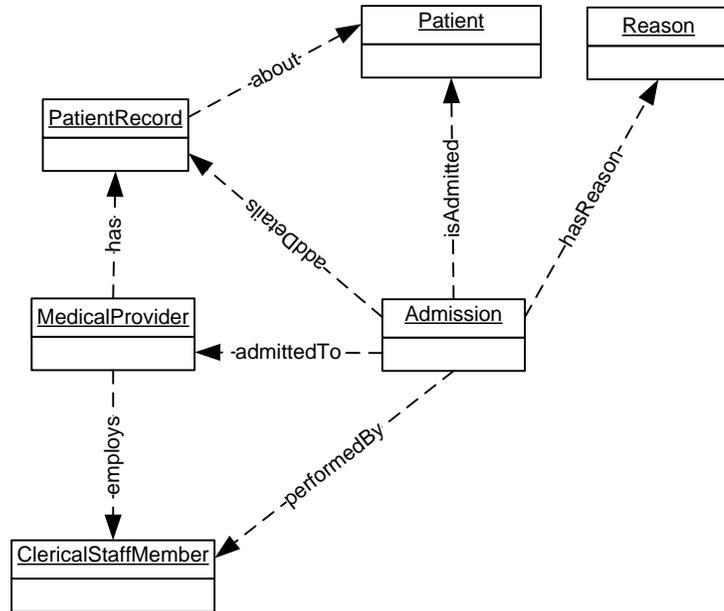


**Figure 24: Medical Ontology - staff concepts**

### 6.2.2. Processes

The previous section has explained various concepts in our restricted medical domain, which will now need to be enhanced with various processes that may occur in this domain. One of the processes would be "admission" (to a hospital). Admission refers to the process of admitting a person for particular reasons to a medical provider, who may or may not have a patient record about this person, whom we refer to as "Patient". During admission, additional personal information is collected in order, for example, to retrieve, amend or correct the patient health record as depicted in Figure 25. The actual process of retrieving the patient record

will not be described here, but would be part of a more extensive elaboration.



**Figure 25: Admission process**

A second process is called "Translation", which can be understood in multiple ways. To our understanding, and use, translation refers to the process whereby a particular entity converts information from one format into another. For example, a translation could be from a real world paper document into its digital equivalent as well as the actual translation from one language into another. This may be needed in the medical domain if a patient is not sufficiently familiar with the language spoken.

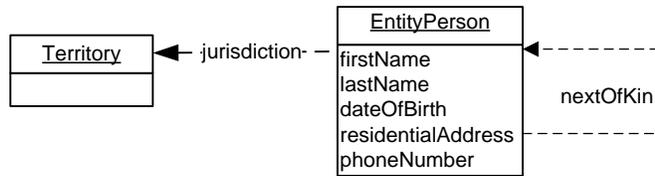
In addition to the processes mentioned above, the medical domain, among others, has processes that deal with personal information slightly differently. Personal information can be generated or obtained indirectly and therefore needs to be assessed differently. However, as they still deal with personal information, we still regard them as privacy processes. An

example of such a process is determining the blood group of a patient. A nurse takes a blood sample from a person and then sends it off to the laboratory to determine the blood group. We regard this as obtaining information indirectly, which affects certain privacy principles negatively, as errors might occur during this process and it may depend on the quality of the blood sample or the lab testing it. Another example would be a doctor assessing the psychological profile of a patient. This would highly depend on the doctor being able to make an accurate assessment which then impacts on the data quality (e.g. being accurate).

### 6.3. Medical Privacy Ontology

Next, we need to create the domain privacy ontology for our restricted medical ontology. Hence, we need to take it concept by concept from the application domain and create new concepts that commit to the appropriate concepts of the generic privacy ontology. Furthermore, the concepts must also take on the privacy relevant attributes of the application domain concept.

Our first concept from the domain ontology is "Person", which we can see as a form of entity in the generic privacy ontology. Thus, we create a new concept "EntityPerson" that commits to all attributes and relationships of "Entity", namely the identifier and the relationship "jurisdiction", which associates the territory with the entity. Second, it takes on the attributes "firstName", "lastName", "dateOfBirth", "residentialAddress" and phone number as well as the relationship "nextOfKin" as depicted in Figure 26.



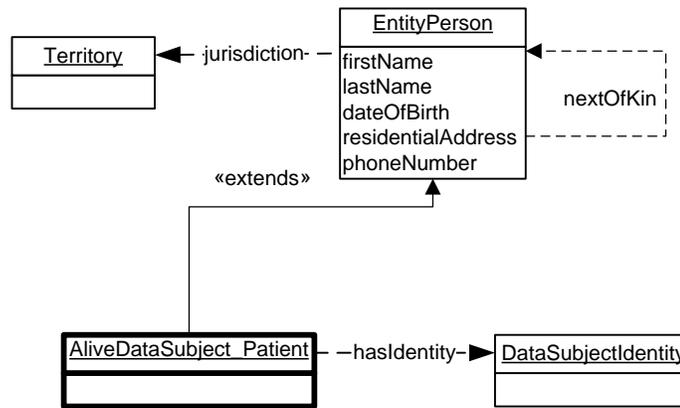
**Figure 26: EntityPerson**

The second concept in our domain ontology is "Patient", which we classify as a "DataSubject" in the generic privacy ontology. The newly created concept in our domain privacy ontology is called "AliveDataSubject-Patient" and commits to all the different attributes and relationships of "AliveDataSubject". It also takes on the privacy relevant attributes of "Patient" as shown in

Figure 27.

Now, the various staff members need to be classified. Generally, they are all classified as entities as a specialization of the previously created concept EntityPerson. Although a hospital has a myriad of staff, we will not bother to classify all of them as it is not relevant to our purposes here. We assume that we have a particular clerical staff, which is mainly a receptionist as well as a doctor who takes care of the patient. In our generic privacy ontology, we classify "Receptionist" as a ResourceHandler, who is permitted to read and "handle" personal information. The concept of this representation is called "ResourceHandler\_Receptionist". As discussed in the previous chapter, a resource handler can - from a privacy perspective - read and "translate" personal information. This includes the receptionist as he or she can take personal information and enter it into the hospital computer system, where we assume that the patient has filled out a paper-based form, which is taken by the receptionist to enter the details. Although the actual permissions on a computer system may require the receptionist to

alter personal information, this is not the case from our privacy ontology perspective, as he or she acts on behalf of the patient only.



**Figure 27: AliveDataSubject\_Patient**

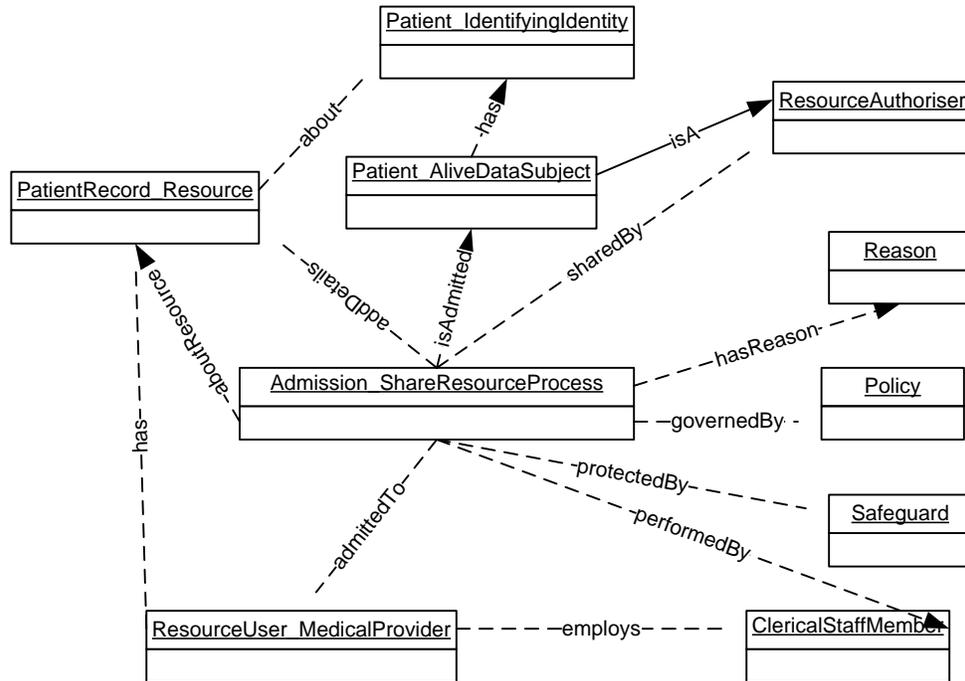
The concept of Doctor can be classified further within the domain ontology, e.g. GP, Dermatologist or Dentist, but the general concept of doctor suffices for our needs. Normally, it is assumed that any medical doctor involved in a patient's care in a hospital has access to personal information of the patient unless a personal conflict exists, which could restrict access to certain details only. In our generic privacy ontology, a doctor is classified as `ResourceModifier` if in charge of a patient and we name the new concept `ResourceModifier_Doctor` in our restricted medical domain ontology.

We consider the most important concept of our restricted medical domain to be the patient record, containing patient information, most of it of a sensitive nature. Naturally, the patient record refers to the concept of resource in our generic privacy ontology, with its individual resource elements. They in turn refer to the various entries on the patient record and need to be classified accordingly. For the sake of clarity, we assume that our patient record has three different types of entries, which are demographic information, medical history and psychological profile. In reality however, patient records will have many more entries in various

other categories, which are most likely cascaded and structured further. However, our limitation here is satisfactory for demonstration purposes without loss of generality. We assume that demographic information is categorized as "IdentifyingResourceElement" and the other two as PseudoAnonymousResourceElements as none has personal identifiers attached. However, this assessment can never be completely precise as one person may have some medical history that is unique across a certain area and therefore has the potential to be identified. Nevertheless, in order to identify or specifically name the person to whom this resource element refers, demographic details - from any source - are necessary. For example, if a patient in a hospital were the only one with a particular illness, others could still not identify him or her without additional information (e.g. name and room number).

Finally, we need to classify the various processes in our application domain ontology. The processes that are relevant to our context of privacy are the ones that are involved with personal information in some way. Hence, any other processes within the application domain are not required to be modeled within the context of our privacy ontology extension. The process of admission is definitely one of them as it involves the collection of personal information from the data subject. Looking at the generic privacy ontology, we can see that a process that involves personal information is called a "PrivacyProcess" about certain resources, which are about a data subject that is performed by a certain EntityIdentity and governed by a certain policy. We can even specify this further as a "ShareResourceProcess" as it involves a resource authoriser, namely the data subject or specifically the patient in our case, and a recipient of personal information, which is a resource user or specifically the medical provider here. We call the new process concept

"Admission\_ShareResourceProcess" in our restricted medical domain privacy ontology as shown in Figure 28.



**Figure 28: Admission\_ShareResourceProcess**

It is noteworthy that the resource authoriser does not need to be the same person as the data subject, such as in the case of minors, where parents provide the necessary information and therefore, are resource authoriser by law.

Now that we have created and classified an example subset of the various concepts in our application domain privacy ontology, we add the mappings to the different quality aspects, namely the privacy principles, which will subsequently help us to evaluate the level of privacy. We decided to show the information ordered by privacy principles instead of concepts as this makes it easier to understand the various implications. Our first quality aspect, data quality, has a weight of "high" in this domain as medical data must be adequate and correct at all times. Otherwise, it could be fatal if

information such as details about allergies were incorrect or missing. As data quality is a function of resource and purpose, we have to consider both concepts together. This is the case during our privacy process "Admission\_ShareResourceProcess". We assign a very high level to all three quality assessment criteria, as the details requested are highly adequate, relevant to the purpose of medical care, and correct as directly provided by the data subject and no other (third) party. Security as the second quality aspect is assigned a weight value of "medium". The reason behind this logic is the nature of this data and the protection of vital interests of the data subject. Therefore, security cannot play such an important role in the medical domain, as methods need to exist that allow employees of the medical provider to gain access to personal information in an emergency situation without explicit consent from the patient. However, due to this, the weight value can be adjusted by the data subject in question to tailor it to their own needs, as they may think that this principle is more important than for an average person. Although we have omitted to mention the actual safeguard concepts for our admission example, we apply a value of 3 out of 5 for adequacy of safeguards and 5 out of 5 for adequacy of personnel requirements. The remaining five other quality assessment criteria for this aspect are not relevant for this process and therefore play no part in the evaluation of the level of privacy and have no impact whatsoever. The data subject's rights are determined by policies that govern privacy processes. Although we have not shown any policies previously as they comprise lengthy documents, we will show a couple of policy concepts for our elaboration. As described in the generic privacy ontology, a policy is a set of statements with certain relationships. In our example here, we assume that the hospital has a statement that permits data subjects to request access to their personal information in writing and patients can view and change it after having made an appointment.

Therefore, we classify the access privileges as "high" (4 out of 5) and the level of ability to request rectifications as "very high" (5 out of 5). Furthermore, the policy also states that the personal information can be withheld for certain purposes, e.g. can be blocked from certain staff members. This is also classified as "very high" for ability to block content for certain purposes. The policy does not state that it is possible to object against processing altogether, which is therefore classified as "not stated" (1 out of 5). We assign a weight level of "high" to this quality aspect as this domain contains highly sensitive information and blocking, objecting and rectifying of personal information or the processing thereof is highly relevant. The privacy principle of "Legitimate Grounds of Processing" is also influenced directly by each privacy process, thus, "Admission" in our example. The aspect of "Unambiguous consent" receives a "very high" value, as does the criteria "Protection of vital interests of data subject". The remaining criteria are ignored for this purpose as they are not relevant. This also applies to the sensitive category criteria, as the whole domain is subject to sensitive personal information, which includes medical data. Therefore, it is not valid to ask if personal details contain medical data as they are the important data in this domain. However, as personal information contain highly sensitive data, the weight is "high" for "Legitimate Grounds of Processing". The "Transparency" principle also receives a "high" weight value as it is deemed highly important in the context of privacy to be notified of personal medical information being collected and processed. Our admission example would map to a value of "before collection" (5 out of 5) for "Data obtained directly, notify of processing" as the patient knows that his data is being collected (as he or she provides it) and will be used for his or her own medical care. All other criteria receive a "very high" (5 out of 5) as well, as the identity of the medical provider is revealed, the purpose of the data is stated and then is

recorded in accordance with current laws, which are at a high level. This has been inferred by the territory in which the medical provider resides. As a third party is not involved, this received a "very high" (5 out of 5) as well. One of the other principles worth mentioning is "Anonymity". We assign a weight of "low" to it as anonymity during the admission process is essentially pointless. The quality assessment criteria of "data anonymized" is specified as "No" and therefore receives a value of 1 out of 5 - however, as the weight for this principle is low, the impact on the overall level is negligible, which makes common sense to have a qualitatively high set of personal information as it is usually vital in the medical domain, and in particular during admission, to know who is being admitted.

Although we have not mentioned every single privacy principle, we have discussed the most important ones and described the impact of every principle. Furthermore, we have shown the various assessment criteria and the values that have been pre-assigned by a domain expert.

#### 6.4. Example of instances

In the previous section, we elaborated the restricted medical privacy ontology on a class level and discussed the mappings to the various privacy principles on a conceptual level. In this section, we will describe an actual example, which can be seen as an instance of the application domain privacy ontology. Naturally, we select the process of admission for our small example as it has been previously discussed at length on a class level.

In our small example, we assume that Lim Kun, a 22-year-old exchange student from China, has to go to his local private hospital in Townsville for a blood test as he is feeling unwell and his body is shaking. On his arrival, he has to fill out the form for his admission, and a new file is created as Lim

has never been to this hospital. The doctor in charge then orders a blood test to check for diabetes. This small example already describes two processes that involve personal information, one being the admission process and the other the process of extracting information about the patient's blood (e.g. type and other substances or the lack thereof in the blood). Now, we consider this example and show how the level of privacy is evaluated.

When going into the hospital, Lim knows that he will be dealing with this medical provider (name: "Park Haven Private Hospital", abn: "31064632613") and the various laws and regulations governing it, which, for example, have implications for the privacy laws under which his information will be legally protected, influencing the privacy principle "Legitimate Grounds of Processing". On arrival, Lim is asked to produce his Medicare card, which he does not have, as he is an exchange student. Hence, he is asked to provide a different form of ID, whereby he chooses his Australian student ID. The student ID contains his full name ("Lim Kun"), the student number ("1432451"), the university at which he is enrolled ("James Cook University"), an image depicting him, an expiry date ("31/03/2010") and a holographic logo for security reasons. A student ID is classified as "IdentifyingIdentity" concept in the demographic domain identifying the student, which impacts on the anonymity privacy principle, inferring that the level of anonymity is low (1 out of 5 stars), but so is the importance of this principle as described in the previous chapter; and hence, this will not have a great impact on the overall level of privacy. As Lim is providing the ID himself, it affects the levels of Transparency and Data Quality, which are both set to "very high" from Lim's point of view - regardless of the actual content of the data. He is asked by the receptionist to fill out the admission form. She is a clerical staff member of the hospital,

whose name "Jane" is displayed on her name badge together with a logo of the hospital. In addition to the details on his student ID, he has to provide his residential address ("12 Derby Street, Mysterton"), his phone number ("0416960612"), his date of birth ("12/03/1987"), his gender ("male") and his next of kin - which we will omit here. In addition to his basic demographic information, he has to provide any relevant medical history ("nothing relevant") and any kind of medication he is currently taking ("none") as well as any allergies ("none") to medications. Finally, he agrees to the policy of the document and gives his consent that the data provided can be used and stored for the current purpose and future visits. The paper form which is the medium for recording Lim's information has no direct safeguards. The only safeguard in place is his ability to fill out the form in a quiet, private area, which impacts upon the principle of security by assigning a "medium" safeguard to this activity. Afterwards, he hands over the form to Jane and she types everything into her computer without any further interaction between them apart from asking him to wait until called. The action of entering Lim's details into the hospital's computer system is a translation process as she converts from one format to another without altering the content semantically as best as possible. This process does not give the receptionist (who reflects the concept of "Staff\_ResourceHandler") any further consent to use his details for any other purpose or to reveal it to anyone else, than Lim has consented to in the first place. Although Lim is not aware of any safeguards in place once the details have been entered, the laws and regulations of the territory provide adequate requirements to medical providers. The hospital computer system is designed to not store personal details with medical data. Every section of medical information gets an arbitrary identifier that can link it with personal details if required. Therefore, the individual pieces are pseudo-anonymized to an extent. After a while, doctor "John" (a Doctor\_ResourceModifier) , who accesses the file

and sees the patient, requests a blood test as he believes Lim may have diabetes. The nurse in question takes a blood sample and affixes an ID that is linked in the computer system to the vial. The pathologist checks the blood sample and enters the details of glucose blood particles into the file for the ID in question. Although this piece of information is about our patient "Lim", it has not been obtained directly from him, but has been assessed by a third party, which could introduce errors and therefore impacts on the principle of Data Quality in terms of impacting on the criteria of correctness, setting it from "very high" to high (4 out of 5 stars). We will conclude our instance example here and continue in the next section evaluating the values that have been collected throughout and describe the level of privacy that is experienced.

## 6.5. Evaluation

We assume that Lim Kun is happy with the standard weight template of the restricted medical domain that has been defined by the domain expert with regards to the level of privacy for the various privacy principles. The weight level of every quality aspect is shown in the table below as we have determined for our example.

<b>Privacy principle / quality aspect</b>	<b>Defined impact weight</b>
<b>Data Quality</b>	high
<b>Security</b>	medium
<b>Data subject's rights</b>	high
<b>Legitimate Grounds of Processing</b>	high
<b>Transparency</b>	high
<b>Consent</b>	medium
<b>Anonymity</b>	low
<b>Finality principle</b>	medium
<b>Processing by a 3rd party</b>	high

**Table 1: Privacy principles weight 1**

This weight table applies to the various processes and concepts across our restricted medical domain and all of our evaluations are based upon them. When calculating them, we convert the weight values to numbers, which are 0.1 for low, 0.5 for medium and 0.9 for high.

The evaluation of the different steps is done in tabular form. Below, we describe the various steps and show the influences in table xyz.

Step 1: Lim enters the hospital. This refers to the activity of consciously choosing a hospital due to a medical problem. This step introduces Lim to the medical domain, which applies the weight factors ( $w$ ) as listed in Table 1 that are used in the privacy level calculations. It also introduces to the concept of medical provider with a certain territory and all its legal implications.

Step 2: This step comprises the admission process, where Lim has to provide his student ID, fill out the form and his details are entered into the hospital computer system. This introduces a number of concepts, which are his identity and his personal information as well as medical details, which are conceptualized as resource elements. Furthermore, the hospital policy is added. In the next part, the concept of receptionist as a staff member of the hospital is introduced who takes the details from the paper-based form and adds them to Lim's newly created personal medical record.

Step 3: In this step, the concept of a medical doctor is introduced, who has access to Lim's personal information and can amend it as required.

Step 4: In this final step of our example, additional details are added indirectly by means of a blood test, which is analyzed by a pathologist.

	Step 1	Step 2	Step 3	Step 4
Data Quality	"n/a"	5	5	4.6
Security	"n/a"	4	"n/a"	4
Data subject's rights	"n/a"	3.75	"n/a"	3.75
Legitimate Grounds of Processing	"n/a"	5	5	5
Transparency	"n/a"	5	"n/a"	5
Consent	"n/a"	5	5	5
Anonymity	"n/a"	1	1	3
Finality principle	"n/a"	3.6	3.6	5
Processing by a 3rd party	"n/a"	"n/a"	"n/a"	"n/a"
Total:	"n/a"	<b>4.48</b>	<b>4.6</b>	<b>4.62</b>

**Table 2: Quality aspects: assignments**

Finally, we have to calculate the overall level of privacy experienced for this example, using the formula as defined previously. The table above does not have a value for every cell as not all privacy principles are influenced by every step; these are marked by the term "n/a". As step 1 does not deal with personal information as such, but only sets up the environment by defining the context and hence the impact factors as shown in Table 1, no privacy evaluation is possible.

As the lowest level of privacy emerged from step 2, we use this minimum as the overall level of privacy experienced in this example, which is 4.48 and is therefore rounded down to 4, for which we assign 4 stars and refer to it as a level of "high".

## 6.6. Conclusion

In this chapter, we have provided an extension of the generic privacy ontology for a restricted medical domain as the application domain. Firstly,

we described the medical domain ontology that has then been used in conjunction with the generic privacy ontology to build the ontology commitment of the restricted medical domain privacy ontology. We have named various concepts, relationships and processes and assigned concrete attributes to them and how and with what impact level they map to the privacy principles. This was followed by an actual example, which provided the instances for our concepts, attributes and processes and concluded with an evaluation thereof, describing a weight impact table as well as the actual level of privacy reflected by that example.

The example in this chapter was primarily focused on the user perspective. In the next chapter, we will provide a more provider-oriented example from the e-commerce domain that shows how the ontology helps e-commerce service providers to integrate privacy related concepts and mechanisms into their system and processes.

## 7. Specialization II - B2C E-Commerce domain

### 7.1. Introduction

In the previous chapter, we described a user perceived example from a restricted medical domain. We have described the concepts of that application domain and their usage to create the application domain privacy ontology that commits to the generic privacy ontology and commits to a number of selected concepts from the domain ontology. We then described an instance example for a hospital scenario. The example was given from the user's perspective, based upon established concepts, processes and procedures within that domain.

In this example, we focus on another application domain, by providing an e-business scenario for a new online shopping system that needs to deal with privacy related issues. The discussion is based upon our previous work which has been quite successful in this domain and published in an IEEE transaction [70]. Essentially, the ontology is used to add privacy concepts into the shopping system while it is being built to allow the engineers of that system to adhere to industry standards and legislation with regards to privacy. It allows the engineers to protect their customers' privacy by integrating relevant mechanisms into it. As engineers are not necessarily experts in the context of privacy, they use this ontology to gain a greater understanding of its concepts and dimensions as well as their influences on the overall level of privacy the customers experience. This is beneficial for both customers and service providers as both can be confident that personal information is protected appropriately and financial and image loss is less likely to occur from loss of personal information.

The example will show a restricted e-commerce domain as it is not possible or useful within the scope of this thesis, to show the myriads of concepts that are available in this domain. The primary focus is therefore on the concepts and processes of an e-commerce business-to-customer (B2C) website, like Amazon. A further limitation applies that we will concentrate on the customers' expected perspective and will not go into details about ordering products from distributors and so on. We will make the assumption that we always have enough products in stock to meet customers' demand. Generally, we will base our discussion upon the work of Chan et al. [71].

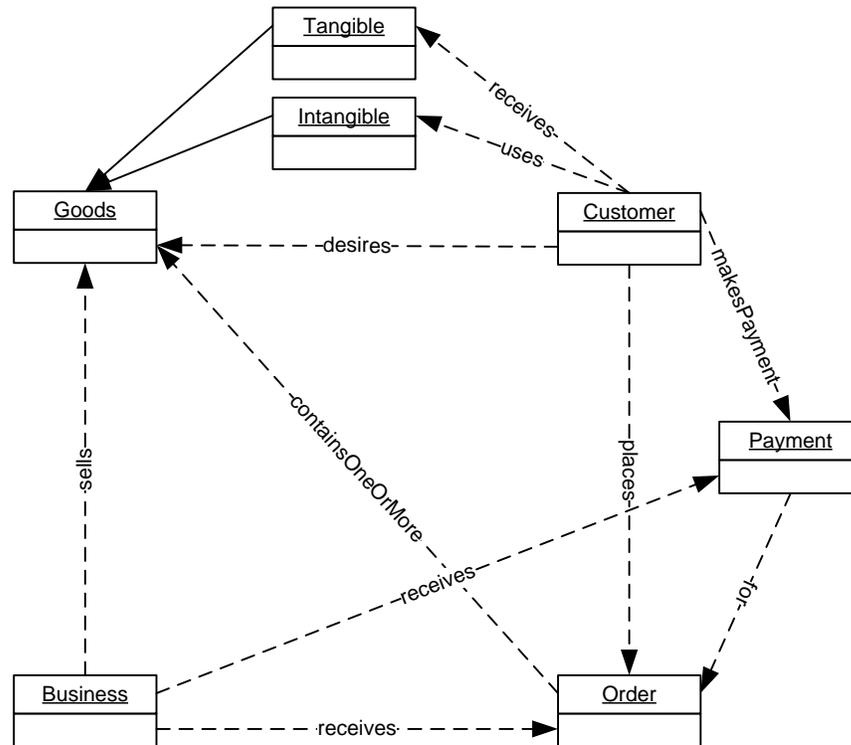
## 7.2. Restricted B2C E-Commerce Domain Ontology

### 7.2.1. Basic high level concepts

On an abstract level, every B2C e-commerce system, just like a normal business in the real world, will have to deal with the concepts of "Goods", "Customer", "Merchant", "Order", "Payment" and "DeliveryMethod" as discussed in [71]. The main difference between a B2C e-commerce website and a real world business is the availability of a shop-front, which is not necessary in e-commerce. As goods are on display and orders can be made online, it reaches a greater potential customer audience by far. Furthermore, the online retailer can provide a search engine and has the ability to personalize the website experience based upon the consumer, which is not available in traditional B2C systems.

The basic concepts and relationships among them are described in Figure 29. Businesses are, by definition meant to sell goods, which can be tangible or intangible. Customers usually desire goods and place an order for them. Afterwards, they have to make a payment for the order and receive the

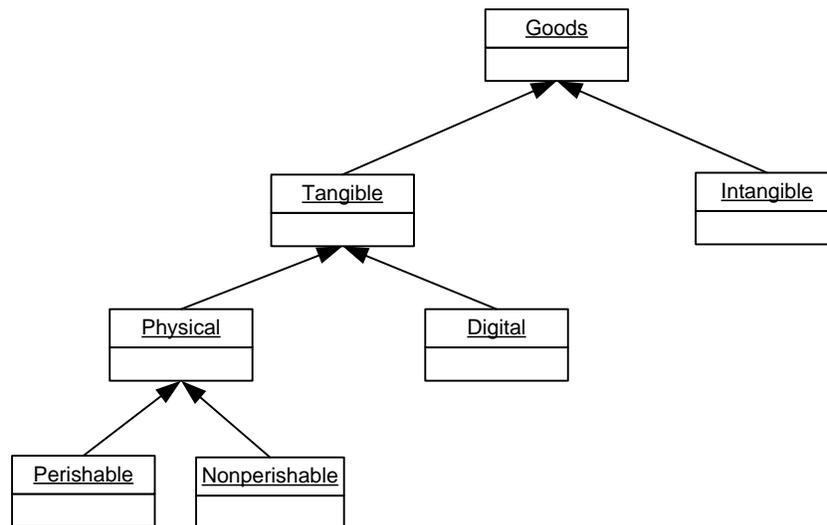
tangible goods. We have decided to distinguish here, as intangible goods cannot be received by definition, but we leave the specification open and define this relationship as "use". This means that an intangible good, also known as service can be used in some way after it has been paid for.



**Figure 29: B2C part overview**

The structure of "Goods" is shown in Figure 30. Goods as the upper or top concept can be specialized further into tangible and intangible goods, where intangible goods are commonly referred to as "Services". Tangible products need to be specialized further into physical and digital goods, as this has certain implications for their storage, distribution and delivery. Finally, physical goods need to be categorized as perishable ones and non-perishable ones, as they have different restrictions for their storage and delivery times that need to be taken into account when modeling (e.g. has attributes that define expiration date and storage requirements such as

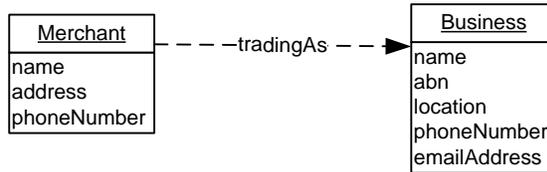
temperature) and physically stocking them. However, as we are not interested in the actual products as they contain very little or no privacy-related data, we will not provide any further details or categories for different goods.



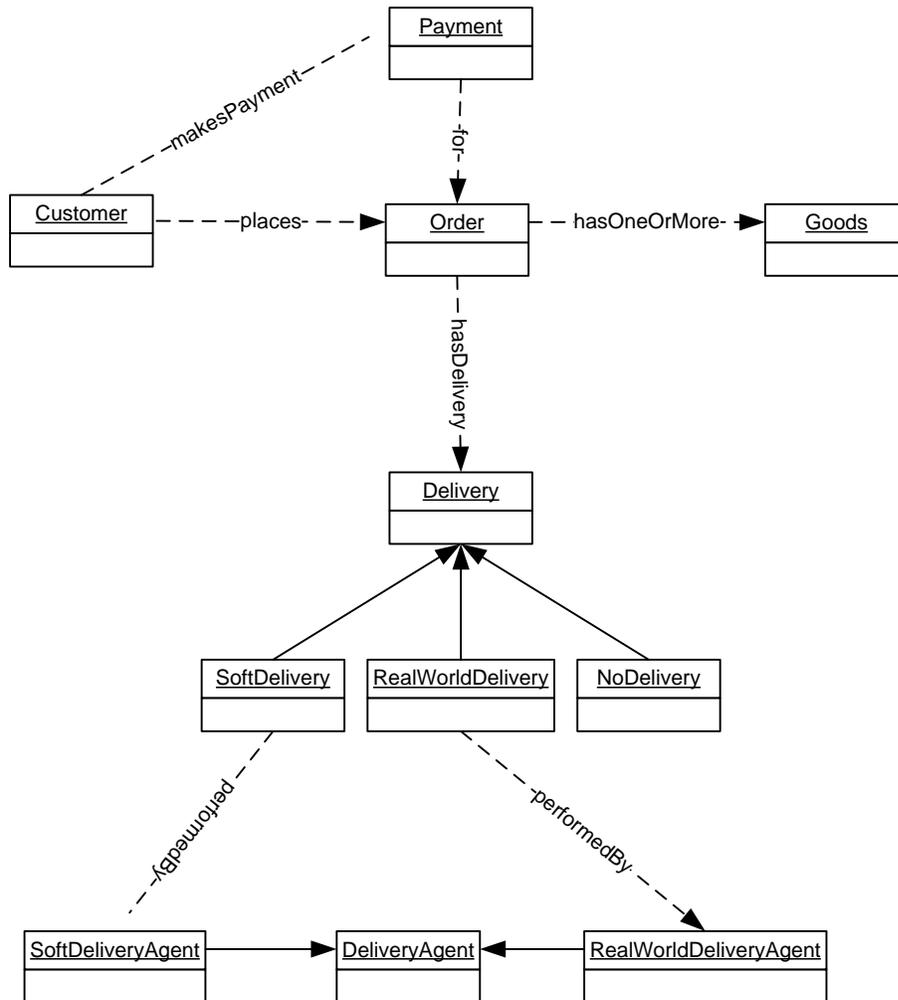
**Figure 30: Goods hierarchy**

In our highly abstract perspective, we also have to distinguish between concepts and processes as well that may be mistaken for each other as they may be commonly referred to by the same name. For example, an "order" could refer to the actual process of ordering something or to the order record or details. Thus, we have decided to not use the term "order" and "delivery" in instances where it may be unclear. Instead, we refer to the "OrderDetails" as the static concept of the actual order record containing all the details of the purchase. The actual process of ordering goods is then referred to as "OrderProcess" with its inherent attributes and relationships. The same applies to "delivery", where we define the static concept "DeliveryMethod" and the process of delivering is named "DeliveryProcess" and elaborated further down the track in this section.

In general, a merchant is either an individual or an organization who trades as a business under a certain business name as shown in Figure 31. This becomes important as it influences the place of jurisdiction, for example, and makes it possible for merchants to trade as multiple businesses.



**Figure 31: Merchant trading as Business**



**Figure 32: Order and Delivery**

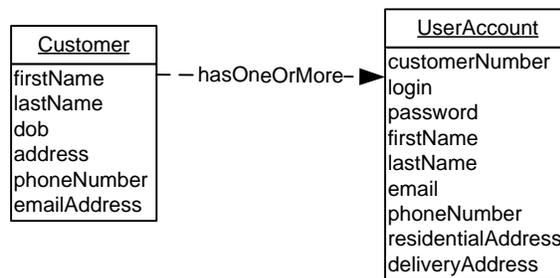
The order that is placed (as in OrderDetails) by the customer contains certain attributes and relationships. It has a relationship with the various goods that have been ordered and with a relationship to a delivery method. Without loss of generality, we assume that only one delivery method can be chosen for each order and that each order is composed of either intangible or tangible goods and may be either physical or digital goods. The actual delivery then has to be performed by a certain category of delivery agent, depending on the type of goods, which can be a real world postal agent or software distribution service provider.

#### 7.2.2. Technical level concepts

In the previous section, we described the various concepts on a business level that is more abstract than a technical level. As our privacy considerations are closely related to the technical concepts, processes and policies, we will now provide the technical level on a conceptual basis. This means that we will provide concepts like shopping cart, but will not delve into their actual technical implementation (e.g. we will not specify details of the kind of software or hardware used or required to implement the technical concepts) as our approach is implementation level independent.

Naturally, we have to model the same or similar concepts as in the discussion above. In the digital world, we assume that individuals - our customers - are represented by a certain account they choose to create. This account includes certain details like a username, password, email address, first and last name as well as residential and delivery address(es) as shown in Figure 33. In addition, it could also include financial data, such as credit card or bank account details. However, best practice demands that this financial information not to be stored with the user account, but that it be used and requested every time a purchase is made. The other

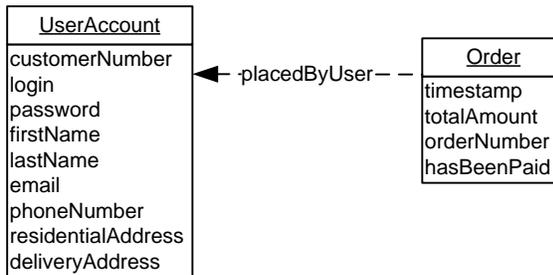
option would be to store financial details separately from the user details and link them to the user account. This could be done in such a way that different access policies and mechanisms apply to financial details than to normal user information. In our model, the user account is the concept that is used throughout the various processes, but it is still the customer who is actually behind this user account and who decides upon the instances of the user account attributes, desires the products and makes the actual purchases and receives the goods.



**Figure 33: Customer and UserAccount**

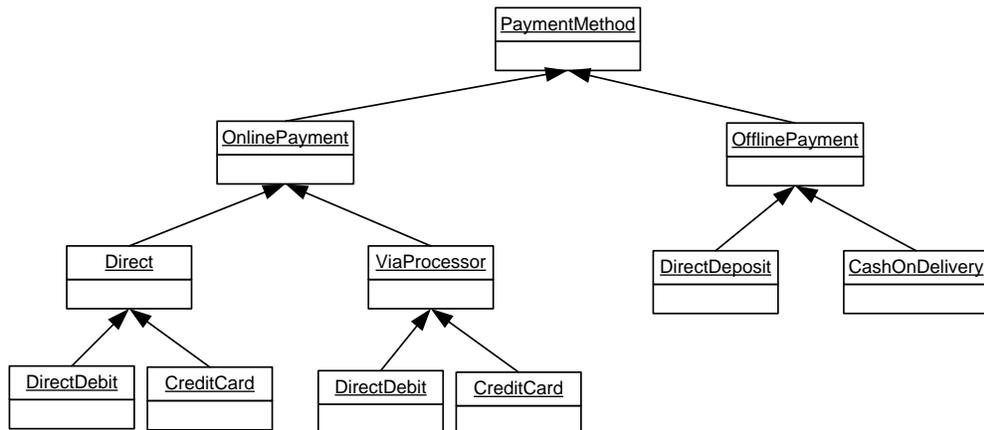
The next concept required in our technical presentation is the actual order (not the process), which is similar to the one in our higher level discussion. For example, an order can contain an order number, a timestamp showing when the order was placed and a total value as well as a boolean value that the order has been paid for. Normally, our details are stored in a database and when conceptualizing our approach, we can keep this in mind to make the transition to a database easier. Therefore, we decided to refer the order back to the user account and not vice versa as in our previous higher level abstraction depicted in Figure 34. The order also has an association with the payment method. Generally, payments can be made via a third party payment processor, or directly. The difference between them is that the former approach does not reveal financial details to the service provider. Furthermore, it may increase a customer's level of trust, as he can

use a certified financial provider (e.g. Paypal) that he may have been used previously and that offers a certain level of protection.



**Figure 34: Order concept**

As we want to show privacy implications here, both options are provided and the outcome of the chosen method will impact on the privacy quality aspects. Without loss of generality, we limit ourselves to payments that have to be made prior to the delivery of the goods, which includes cash on delivery (COD) as purchase orders and tax invoices are often not offered to end customers (in a B2C conceptualization) and have other implications that are beyond the scope of this work.



**Figure 35: PaymentMethod**

Payments can be classified as online and offline payments, which have implications for certain privacy dimensions as discussed below. Offline payments refer to payments made in way other than via the merchant's

website. This can include online payments (e.g. direct deposit) via the website of the customer's bank. However, as this is beyond the scope of the merchant, it cannot be explored any further here. Online payments, on the other hand, are made directly after the order has been placed, either by entering financial details on the merchant's website, who processes them directly, or via a payment processor. The methods available to the customer are determined by the payment processor and therefore - again - are beyond the scope of the merchant and hence, our discussion. However, the merchant does receive a confirmation from the payment processor that a payment has been processed successfully, allowing the merchant to proceed with the order process. Direct payments made via the merchant's website can include credit cards and direct deposit payments where customers have to enter their credit card details or bank account details. The merchant has the option to save those details for subsequent orders, which, however, influences some privacy dimensions, depending on the storage facilities available.

### 7.2.3. Technical level processes

After discussing the main concepts in the context of privacy, we elaborate the various processes and workflows that use personal details from the initial contact with the customer to the finalization of an order.

The initial contact with a customer is the visit of the e-commerce website. Customers usually do not need to register in order to visit an e-commerce website and browse the goods advertised. Customer visits to a website reveal a small portion of semi-personal information that includes IP addresses, which can imply other details about the customer's country and internet service provider if not obfuscated in some way by the customer, which, however, is not known to the other party. It also contains details

about previous visits (in the form of cookies, for example) or web browser information. The information revealed does not necessarily identify a single person, but a computer in many cases that may be used by multiple persons at the same time. However, for our elaboration we will not pursue this avenue any further within the scope of this thesis.

The initial process is the registration of the customer where he creates a user account that contains details as described above. The account is assigned an automatic user account number and the user has to choose a username and a password. The length and complexity of the password protects the user account accordingly. The other details are all user-chosen and are not verified by the merchant in general; however, some require a verification of the customer's email address. By registering, customers also agree to the terms and conditions of the website / business and may be asked if their details can be used for promotional purposes like email advertisements from the business or its partners. From the time of registration, subsequent website visits can be tracked to particular user accounts, and hence, to the associated real world customer if the account information is valid.

The second process is the process of ordering goods from the business, which involves the selection of the actual goods and their quantity. During the checkout process, the customer will be asked to provide additional personal information if this is different from the one's registered. This may include different shipping addresses for example. The checkout process is usually completed by a selection of the payment and delivery methods as well as the user agreeing to certain additional terms and conditions. We have opted to show details of the payment in a separate process, although

this may be integrated into the actual ordering process. The order is usually acknowledged by an email to the customer's email address.

The third process is about payment for the goods ordered. The customer has a choice of direct payments or payments via a processor as discussed above and has to provide financial details as required, which may be verified immediately (e.g. credit card details) or require a certain amount of time to process (e.g. direct debits).

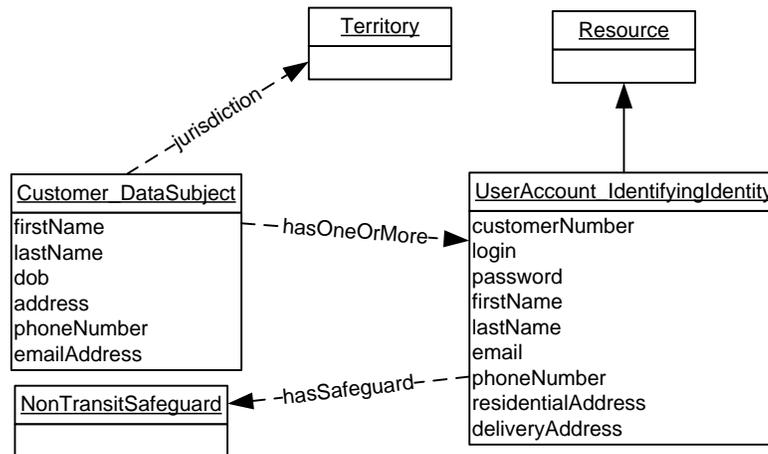
The fourth and final process is dedicated to the delivery of goods. We omit the concepts of waiting for stock to arrive in the warehouse and assume, for the purpose of our discussion, that stock is available at all times. Depending on the type of goods, delivery may be instant (e.g. a download) or may have to be passed on to a delivery agency. We assume that pick-ups are not possible in our example. As downloads are a simple process, a real world delivery involves additional parties with which certain personal information needs to be shared, such as name and delivery address.

### 7.3. E-Commerce Privacy Ontology

After we have described the various concepts, attributes, relationships and processes in this application domain, we continue by creating the restricted B2C e-commerce privacy ontology. This ontology is created by using the concepts from the application domain ontology and the relevant generic privacy ontology concepts in order to create the new application domain privacy ontology concepts and relationships.

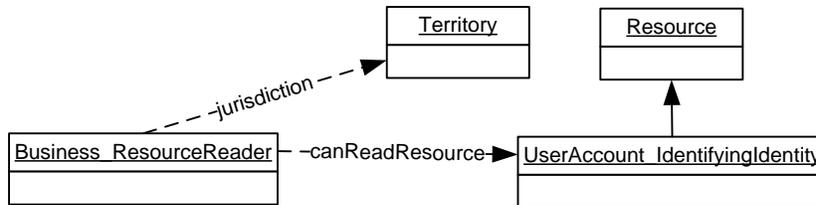
The first two of our concepts are the customer and their user account. We have identified them as the concept of "Customer\_DataSubject" and "UserAccount\_Identity" respectively. We name the user account identity as

a customer can have multiple at any given time and can very freely decide on the instances of the concept attributes. As identities are merely resources as depicted in Figure 16, identities require safeguards to protect the concept from unauthorized access. We assume that the identity identifies the user as he or she is likely to be wanting to receive the ordered goods as shown in Figure 36. However, the impact of false information is marginal for the business until the customer actually places an order. The only information that can be ascertained is that the customer must have had access to the email address, as it had to be verified before activation.



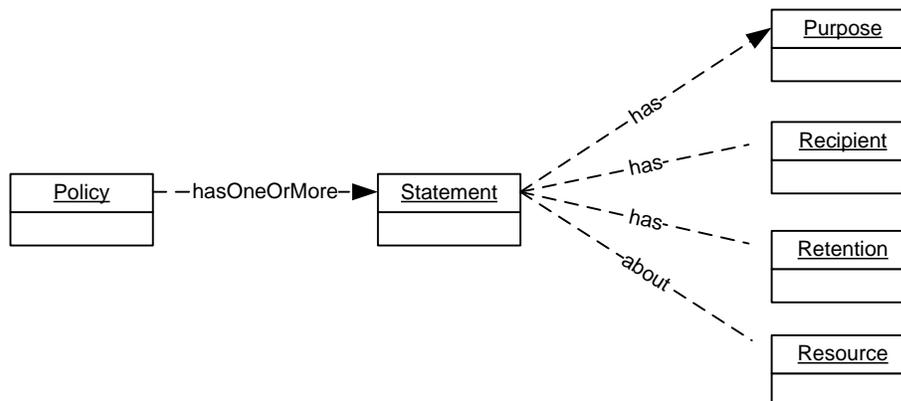
**Figure 36: Customer\_DataSubject and one of its identities**

Secondly, we model the concept of business from the application domain. The newly created concept is named "Business\_ResourceReader", as the business is an entity that can access (but not modify) personal information and hence can be classified as "ResourceReader" as shown in Figure 37. One of the resources to which the business will have access to is the user identity.



**Figure 37: BusinessResourceReader**

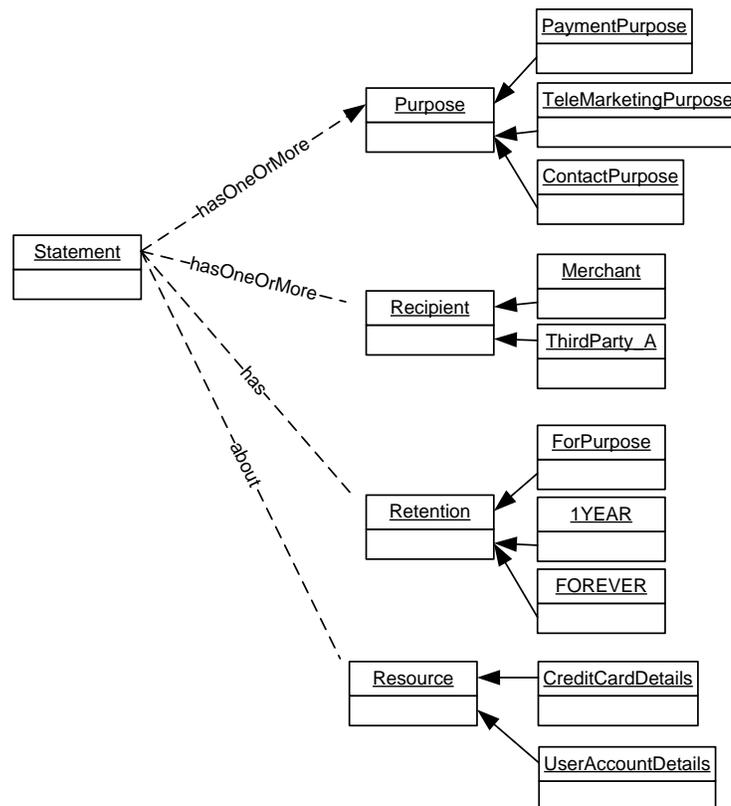
The third important concept is the terms and conditions of the website and its usage. Essentially, it is a policy with various statements that describe how personal information is stored, used and accessed. An example of these statements is shown in Figure 38 and more specifically in Figure 39. However, the actual selection of policies can only be determined at instance level as it is up to the merchant or business to determine the content, which can be selected from available statements in the application domain privacy ontology.



**Figure 38: Policies and statements**

Figure 39 shows an example of building blocks for our e-commerce description. Essentially, the system developers will need to select appropriate classes for purposes, recipients, retention and resource in order to assemble the statement. Instance-wise, they need to add the attribute instance for these concepts, but for clarity, we have omitted the attributes in this figure. For example, one could decide to build a statement

that is about the credit card details of the user and has as its purpose the "PaymentPurpose". Furthermore, it may have a recipient concept of "Merchant" and a retention of "Forever". When choosing these blocks, it is immediately clear how the quality aspects are influenced as the mapping is provided (see below). For instance, the concept of retention influences the privacy principle of "Finality Principle" in the generic privacy ontology and choosing a retention time of one year provides a concrete mapping to the quality aspect of "retention period", to which the domain expert has assigned a level of "Specified period - not anonymized", which evaluates to 3 out of 5 stars and, hence, has a negative impact on the overall level of privacy. Therefore, by selecting the retention concept of "ForPurpose", this would change this impact level from 3 stars to 5 out of 5 stars, thereby improving upon the overall level of privacy.



**Figure 39: Policy statement template in the e-commerce domain**

Before we continue with the actual instances of our example, we need to define the weight level in this domain that is applied to the quality aspects of our evaluation. We cannot simply take the weight factors from our example in the previous chapter and re-use it here, as the requirements and type of data concepts have changed. For example, instead of medical data, we have to deal with financial information that requires different privacy preservation principles.

<b>Privacy principle / quality aspect</b>	<b>Defined impact weight</b>
<b>Data Quality</b>	medium
<b>Security</b>	high
<b>Data subject's rights</b>	medium
<b>Legitimate Grounds of Processing</b>	high
<b>Transparency</b>	high
<b>Consent</b>	high
<b>Anonymity</b>	medium
<b>Finality principle</b>	high
<b>Processing by a 3rd party</b>	high

**Table 3: Weight / impact values in the e-commerce domain**

In our restricted B2C e-commerce domain, we have selected the weight values as follows: "Data quality" that deals with correctness and adequacy of information, is set to a "medium" impact factor, as incorrect information, for example, will have no real damaging fact as it would in the medical domain. This is easy to understand as incorrect medical details may have life-threatening consequences, while this does not apply to either demographic or financial details, remembering that this is from the perspective of the customer. "Security", as the second quality aspect, has a "high" impact factor as it is necessary to protect financial information with a high adequacy level of safeguards as unauthorized access to financial details can produce financial losses for the customer. "Data subject's rights", as the third quality aspect, has a "medium" impact factor in this domain as there is no immediate requirement to amend or alter personal details and the consequences of delayed updates and access are

comparatively lower. "Legitimate Grounds of Processing", as the fourth impact factor, has a "high" impact, as the entity (the business) processing personal information must have legal grounds for processing personal information (i.e. it must be a business) and unambiguous consent from the data subject to process it. Any activity contrary to that could potentially have a significant impact on customer-experienced privacy. "Transparency", as the fifth quality aspect, has a "high" impact factor weight, as it is essential for the customer to be notified about any collection of personal information and also about who has access to what personal information and for what purpose, especially when dealing with financial information. "Consent", as the sixth quality aspect, has a "high" impact, as it is vital for the data subject that the processing party uses personal information for the agreed purpose only, as failure to do so could lead to personal details being revealed to others, which is not desirable in this domain. "Anonymity", as the seventh quality aspect, has a "medium" impact factor since anonymity is not possible when ordering goods online and having them delivered. However, personal details should be anonymized when they are not used for the actual purchase of the order but for other organizational tasks. The "Finality principle", as the "eighth" quality aspect, has an impact factor of "high" as personal information, especially financial details, should be used only for the purchasing purpose and not be kept for any other reason. "Processing by a third party", as the ninth quality aspect, has a weight impact factor of "high", as personal information should not be revealed to other parties unless necessary for the fulfillment of the contract, for example delivery (i.e. the postal agency will need the name and address for delivery). However, other personal details, in particular financial information, should not be shared with other third parties that may not even be known to the data subject, without unambiguous consent. This is often the case in this domain, as personal

information (e.g. what kinds of products is a certain customer interested in) are sold for telemarketing purposes, for example.

After we have defined the impact factor of the various privacy principles, it is necessary to define which concept and process impacts upon which principle. However, we are discussing this in the next section since our instance example will have to select various concepts and their associated implications in order to assess the level of privacy that is best for the customer. Therefore, we want to show the various options available and choose to do this on an instance level.

#### 7.4. Example of instances

In our instance example, we assume the position of developers of an e-commerce B2C system that is being designed and built and has the goal of setting up an online presence for customer interaction and purchase (i.e. a shopping website). The primary goal of our approach is the integration of privacy related concepts and principles to design it from the ground up with customer privacy in mind.

##### 7.4.1. Concepts

Selecting the types of information that are required to register and order a product is essential in order to determine whether they are not excessive. The developers can select various resource element concepts that form part of the identity, which in turn reflects the concept of UserAccount. The attributes chosen for the UserAccount have been shown in the previous section and are a customer number (e.g. "123456"), a username (e.g. "johnsmith"), the password, which is a highly sensitive resource and therefore should be protected by the concept NonTransitSafeguard called

"MD5Hash" (which hashes the password according to the MD5 algorithm and then only the hash is stored, but never the plain-text password) to achieve a higher level of privacy as described below, the first name ("e.g. "John"), the last name (e.g. "Smith"), the email-address (e.g. "johnsmith@email.com"), the customer's phone number (e.g. "+61456789012"), the residential address (e.g. "1 George Street, Sydney, NSW 2000") as well as a delivery address (e.g. "1 George Street, Sydney, NSW 2000"). The selection of appropriate passwords by the customer also impacts upon the level of privacy, since weak passwords are prone to brute-force or dictionary attacks. However, the actual strength of a password is not evaluated with the concept of a user account, but during the process of registration or the process of password change, where certain policies can be applied that enforce complex and lengthy passwords which in turn impacts on the relevant quality aspect of "Security". Hence, the password itself is a safeguard ("access control") as well as being protected by a safeguard (a non transit safeguard). The "DataQuality" quality aspect that assesses the adequacy, relevance and correctness of a resource (i.e. personal information), has an assigned value of "very high", as all the information as mentioned above is relevant to the process of ordering a product. Furthermore, the data subject (i.e. customer) provides the actual information and, hence, they it can be deemed to be correct. In general, no party other than the customer would alter personal account details here, making the data subject the sole "ResourceModifier" for personal account details (which implies the actual very high level on correctness of data quality). However, the customer can advise, hence authorize the business to make alterations on his behalf, for example in the case where accessing the data is not possible. This would not alter this assessment and hence the quality aspect. Security as another privacy principle is directly impacted upon by the quality of protection in terms of

storage for the whole repository of personal information. We have already mentioned the safeguard for the password, which is stored in its hashed form, not in plain-text. This impacts the quality aspect of security at a level of "very high" for the assessment criteria of adequacy, while a plain-text password would score a level of "low" here. Therefore, it is in the interests of the designers to choose the hashing safeguard for the password instead of storing it in plain-text. The second type of safeguard applies to the storage of personal details in whatever repository it is going to be stored (e.g. database). As this type of personal information is classified as demographic details, it has lower requirements in terms of storage safeguards. Therefore, selecting the safeguard concept of "NoSafeguard" is adequate in this context with the exception of the password, which we discussed previously. The third safeguard refers to the access control of personal information, which also requires an adequate safeguard. The safeguard could be "NoSafeguard", meaning that anybody who has access to the system can access those details. The concept of access safeguard however, is not related to the purpose of the data as this is a different issue described below. We have selected the access safeguard of "EmployedSalesAndSupportStaff" who are permitted to access these details, which is deemed as a "high" safeguard in terms of adequacy. This safeguard would prevent other users (e.g. other customers) from accessing personal information. As the password has an additional safeguard associated with it, it would not be revealed by this safeguard in its plain-text form, but still in its encrypted form, which bears a relatively low risk. A subset of these user account details is created when the product is a physical, tangible good that requires shipping. In that case, the name, delivery address and phone number are extracted and shared with the delivery agency, which is a process that is explained below.

The second concept that deals with personal information is the concept of "Order", depending on how it is modeled in the system. We assume that the concept of order is composed of an order number that is generated automatically (e.g. "10001"), a timestamp that defines when the order was placed (e.g. "01/01/2009-15:31:12"), a boolean flag defining if the order has been paid for and the total value of the order (e.g. "\$AU 277.31"). Furthermore, the order has a relationship with both the customer who has ordered it and with the payment method. This type of concept has no real personal information associated with it apart from the identifier that links it to the user account concept, which makes this a "PseudoAnonymousResourceElement". Therefore, we apply the same safeguards as before, having the same impact on the "Security" quality aspect.

The third concept dealing with personal data is referred to as "PaymentDetails" and belongs to a highly sensitive category as it contains financial details. We assume that both credit card details and bank account details for direct debit have the same level of volatility and therefore need to be protected in the same way. "PaymentDetails" can either be a direct or an indirect concept as discussed in the previous sections, as it can be done via a third party service provider (the payment processor) or directly. The indirect one would contain attributes such as order number (e.g. "10001") as well as a reference number assigned by the payment processor (e.g. "123456789") that would appear on the bank transaction statement of the merchant. For the sake of simplicity, we assume that all financial are transactions carried directly, not via a payment processor. Therefore, all financial details have to be collected from the customer directly. We show how a credit card resource would be used and protected and leave the direct debit example aside as it would work out similarly.

A "CreditCardPaymentDetails" concept would have attributes such as the credit card number, an expiry date, the name of the cardholder and the verification number as well as the type (e.g. "Mastercard" or "VISA"). This resource requires both a "NonTransitSafeguard" as well as an "AccessControlSafeguard" of a greater protection level than the one for demographic details. If we were to apply the same one as previously, the impact level of the assessment criteria "adequacy" for the "Security" quality aspect would decline dramatically. Therefore, we have chosen the access control safeguard of "ProcurementStaff", which allows persons directly involved with the processing of the order to access these details only. Furthermore, we have chosen the safeguard of "AESEncryption" for all details when stored permanently, which may be the case as processing the payment may not be instant due to multiple factors such as unavailability of the internal payment facility. The password for this encryption would be known by members of procurement staff only, or be integrated into their authorization system. We can apply similar rules to the quality of the password as we did earlier with the customer password in order to achieve a higher level of privacy preservation through security mechanisms.

#### 7.4.2. Processes

We have already explained the various processes in this system, mainly the "UserRegistration", "OrderProcess", "PaymentProcess" and "DeliveryProcess". As all of them are involved in dealing with personal information, they are to be classified as "PrivacyProcess" that has certain attributes and relationships. The main relationship here is the association with a safeguard as every "PrivacyProcess" needs to be safeguarded adequately. The "UserRegistration" process involves the user entering his

details online and agreeing to the policies and terms. The safeguard in place would be a "TransitSafeguard" that protects the information end-to-end between the physical entering and the storage. The most common "TransitSafeguard" used nowadays is "TransportLayerSecurity" (TLS) a successor of Socket Layer Security (SSL), which creates an encrypted and authenticated tunnel between the user and the e-commerce system / website (or whatever interface it is). The attributes and details of that safeguard are beyond the scope of this thesis and can be found in the relevant documents (e.g. RFC 2246). This safeguard is best practice and commonly used and therefore assigned a "very high" level of adequacy for "Security". During registration, the user would have to choose a password, which is bound by certain policies that determine the complexity needs. In general, e-commerce websites have to balance between their users' need for strong passwords and their business needs since they may lose customers if the requirements are too high and the user decides not to register, thereby perhaps losing the user's custom if he proceeded with an order. Therefore, in our policy statement that defines the password complexity, we choose a moderate one such as "MinimumEightCharacters", which has a medium level of adequacy for security and hence lowers the overall level, but this is acceptable due to the potential loss of custom if very strong ones were enforced.

The policies and terms of services are a vital part of the user registration and build the contract between the customer and the business when the customer agrees to them - even more when ordering actual goods. We have shown building blocks and excerpts from the policy design in the previous section and need to decide on some of the statements as an example of how such a policy would be assembled and its impact on the level of privacy. As a policy is composed of a variety of statements, we have

to limit ourselves to a very small example in our discussion here. For our example, we use the statement about the financial details. We choose to use the purpose "PaymentPurpose", the recipient concept of "Merchant" and a retention of "ForPurpose". This assigns the best possible values for "Transparency", "DataQuality" and "FinalityPrinciple". If we had chosen a retention of "Forever", the finality principle would have suffered and lower values assigned. We would continue with all other statements in this fashion.

The process of placing an order and entering payment details is similar and has similar safeguards as before. However, the impact of a weak safeguard for financial details is much higher than for demographic details.

The final process in our elaboration is the one that refers to sharing of personal entities. This can include contractors as well as other partners who may have a commercial interest in those details (e.g. marketing / advertising companies). As we assume that our customer privacy is of the uttermost importance, we assume that sharing of personal details is permitted only if unambiguous consent is provided or it is required in order to fulfill the contract with the customer (e.g. for delivery).

We do not discuss the technical implementation of our approach, although this would impact on the levels of privacy. For example, if we had chosen an implementation that had security problems or programming errors and therefore could lead to exposure of personal information, this would assign lower levels for the relevant quality aspect on an implementational level evaluation.

## 7.5. Evaluation of the abstract technical level

	<b>C1</b>	<b>C2</b>	<b>P1</b>	<b>P2</b>
Data Quality	5	5	5	5
Security	4.5	5	5	5
Data subject's rights	5	5	"n/a"	"n/a"
Legitimate Grounds of Processing	5	5	5	5
Transparency	5	5	5	5
Consent	5	5	5	5
Anonymity	1	3	3	3
Finality principle	5	5	"n/a"	"n/a"
Processing by a 3rd party	5	5	5	5
<b>Total:</b>	<b>4.11</b>	<b>4.29</b>	<b>4.82</b>	<b>4.82</b>

**Table 4: Evaluation of the B2C privacy ontology example**

In this evaluation, we aggregate the concepts of "UserAccount" and "Order" as they are similar with regards to their privacy requirements and refer to them as "C1" here. The payment details are referred to as "C2". The process of registration, which includes the policy, is referred to as "P1" and the process of ordering as "P2". This should suffice in our example here to provide an understanding of the evaluation of the overall level of privacy achieved as we have explained it at length in the previous sections.

Data quality has very high levels (5 out of 5) for all of its assessment criteria (adequacy, relevance and correctness) and therefore, it equates to 5 according to the formula in section 5.7. In our example, security is assessed by the adequacy of its safeguards and its security policies in place, which equates to 4.5 due to the fact that the lowest adequacy for the safeguard is 4 (the one for the password has a medium impact level (3 out of 5)). The data subject's rights, which are assessed by the access privileges to one's own data, the level of ability to request changes and the ability to object against processing, is set to very high (5 out of 5) as well. The remaining

quality aspects all have very high levels, apart from the anonymity one as it is not possible in this domain to order goods and staying anonymous at the same time. However, as the weight factor of anonymity is medium only, its impact is reduced for the overall level. The order process as such is pseudo-anonymous, as the order is stored under the customer number and that requires additional access. Combining the two resources of "UserAccount" and "OrderDetails" lowers the level of anonymity back to "no" (1 out of 5 stars) however. As we evaluate by using the minimum level on the various evaluation steps, we can assume that a medium level for the "OrderDetails" and "no" for the "UserAccount" suffices.

Overall, the level of privacy is evaluated as 4.11 (the minimum of all total values), as shown in Table 4, which equates to a level of 4 stars out of 5 stars and is therefore deemed "high". This is mainly due to the lack of anonymity, which is, however, difficult to achieve.

## 7.6. Conclusion

In this chapter, we have provided an example from a restricted B2C e-commerce domain, whereas developers designing a system need to build it in the context of privacy. Firstly, an application domain ontology was created, followed by the ontology commitment that commits to the generic privacy ontology as well as to some of the concepts, attributes and relationships of the application domain ontology. This has then been evaluated to achieve a high level of privacy for the customer experience. This has shown us that the ontology can aid system and application developers to gain a greater understanding of privacy and assist with their choice of concepts and methodology.

## 8. Implementation

### 8.1. Introduction

In this chapter, we will discuss the choice of language and the tools we have used to create the ontology according to the methodologies we have chosen. Therefore, we discuss RDFS and OWL as languages for the semantic web and elaborate our choices. This is followed by the explanation of our main development tool, Protégé, and we will show some of the concepts and relationships in the form of screenshots with accompanying explanations.

### 8.2. Choice of languages

#### 8.2.1. RDF and RDFS

The Resource Description Framework (RDF) is a language for representing information about resources on the World Wide Web [72]. It is used to make statements such as subject-predicate-object from the English language available in a machine readable format. Therefore, it is one of the choices for our requirements here. RDF has no vocabulary as such and anything can be expressed in it, as long as it follows the principles above. However, as the lack of vocabulary makes it unsuitable for machine processing, as any software would not understand the context or its meaning whatsoever, RDF Schema (RDFS) has been created. RDFS describes a way to define vocabularies that can be used by RDF. With the help of such a vocabulary, which has to be known by all parties using these statements, a principle way of exchanging information and making meaningful statements is possible. RDFS introduces the concept of classes, which allows classes to be

specialized into subclasses that can inherit from multiple upper classes. RDF, on the other hand, is then used to create and use instances of these classes and make the actual statements. On top of the class hierarchy, RDFS introduces certain other types. One of the types is "range", which describes that a certain subject instance has a valid range of objects from an instance of certain classes. Other properties of RDFS include the ability to specify cardinality constraints on properties, such as a child having exactly one mother, or specifying synonyms, such as different classes with different names are actually the same concept. Furthermore, it is possible to create unions and intersections of classes and hence, create new classes thereof. However, ontologies require a far greater vocabulary as well as formal semantics, which is why RDF and RDFS are not sufficient for our needs to create an ontology. However, RDF and RDFS are reused on OWL, which is our language of choice to represent our ontology.

### 8.2.2. OWL

OWL is the web ontology language, a recommendation by the World Wide Web consortium [73]. It has been derived from the previous approaches of DAML [74] and OIL [75] and its successor DAML+OIL [76]. OWL is based upon RDF, but unlike RDF and RDFS, has a far greater expressiveness, as it specifies more than just relationships between resources and a generalization hierarchy respectively and also specifies formal semantics that are required for an ontology language. OWL comes in different "flavours", which are OWL Lite, OWL DL and OWL Full. The main difference between the three is their increasing expressiveness from simple constraints via computational completeness to syntactic freedom of RDF with no computational guarantees [73]. Hence, OWL-Full is an extension of OWL-DL and OWL-

DL an extension of OWL-Lite, or an OWL Lite ontology is also a legal OWL-DL Ontology and every legal OWL-DL Ontology is also a legal OWL Full ontology.

OWL-Lite is used in situations where the features of RDFS (a classification hierarchy) are required in conjunction with simple constraints [77]. As OWL-Lite has the lowest formal complexity (e.g. allows cardinalities of 0 and 1 only), it allows an easier migration path from simple hierarchies such as taxonomies. Furthermore, providing support tools for OWL-Lite is simpler than for the remaining two flavours of OWL due its low complexity. However, OWL-Lite is too limited for our requirements since we require cardinalities other than just 0 and 1.

OWL-DL as an extension of OWL-Lite has greater semantic expressiveness, while retain computational guarantees. This means that it has the properties of computational completeness and a guarantees to compute in finite time, which also refers to it as being deterministic. The DL part of OWL-DL stands for Description Logic, which is a field of study concerned with the idea of automatic reasoning. Automatic reasoning is one of the requirements for our ontology and therefore, we decided to choose OWL-DL. OWL Full is not suitable for our requirements, although it provides greater semantic expressiveness, it lacks computation guarantees and therefore is not appropriate for our needs, as we need to be certain that our ontologies are computable in finite time in order to be useful in their application. We omit further descriptions of OWL, as countless examples and elaborations can be found in other works such as [73].

### 8.3. Tools

#### 8.3.1. Protégé

We decided to use Protégé, a tool for developing ontologies and knowledge-base frameworks from the University of Stanford, as our primary means to represent our conceptual ontology in machine-readable format. It has an easy learning curve, while providing great user support and the ability to export into languages such as RDF, RDFS, XML schema and OWL, our choice of implementation language.

For our purposes of representing our conceptual framework on an implementational level, Protégé is ideal due to its rapid prototyping features. Furthermore, we have chosen version 4 of Protégé, specifically Protégé-OWL due to its built-in reasoner and support for OWL2.0 features.

### 8.4. Examples (screenshots and code snippets)

#### 8.4.1. Main concepts

In this section, we will demonstrate some of the main concepts of our ontology, depicted as screenshots and code snippets from Protégé. We use the built-in OWL Viz plugin, which is based upon Graphviz version 2.2. However, this plugin provides a class level view only and does not provide views for relationships (object properties) and data properties. Therefore, code snippets for data and object properties are provided as well. We have opted to show the inferred representations only as this provides a greater level of detail and relationships in some cases. It can also be seen as a formal validation method to check for inconsistencies in the ontology implementation; however, its intended

use is the inference of the class hierarchy, which shows up errors in cases where a relationship, expression or restriction cannot be satisfied and therefore is invalid. Our main concept hierarchy is the entity hierarchy as depicted in Figure 40.

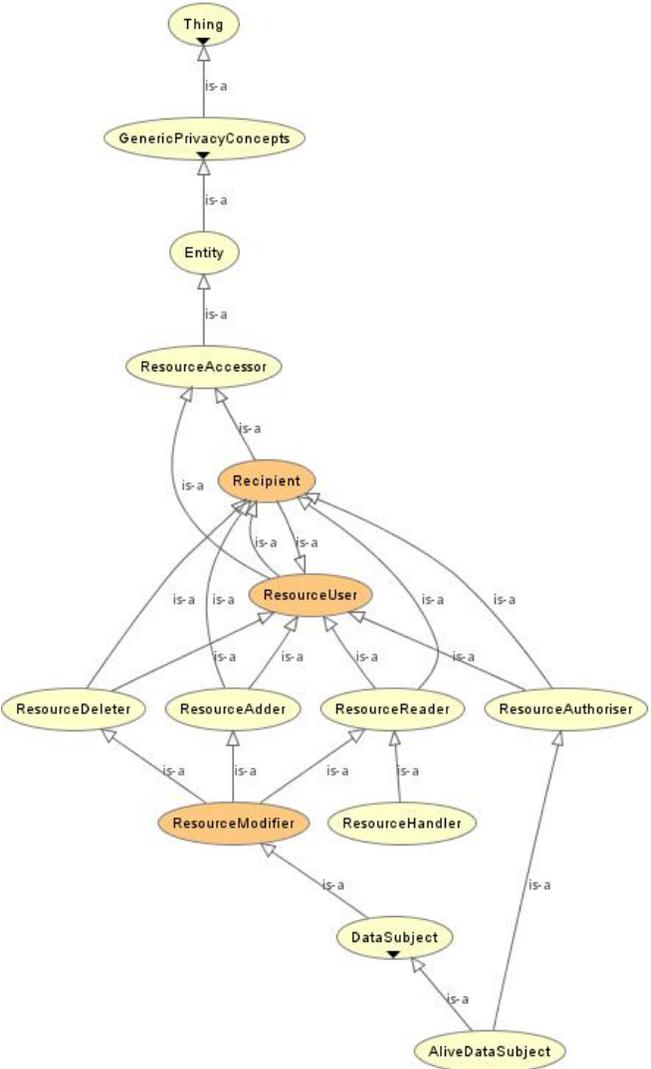


Figure 40: Entity concepts

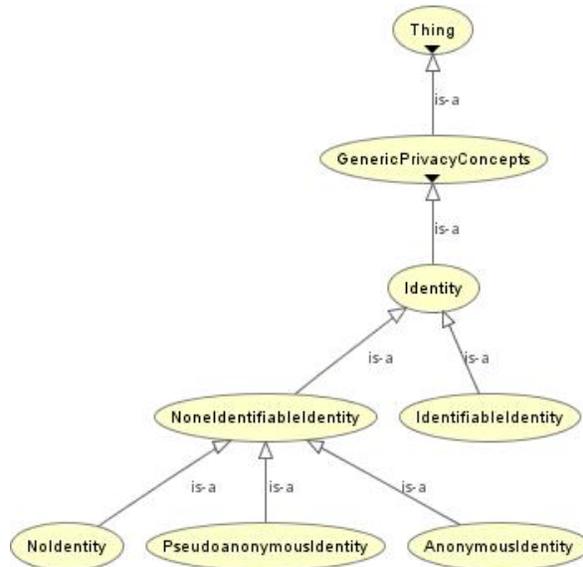


Figure 41: Identities hierarchy

Class Annotations | Class Usage

Usage: NonIdentifiableIdentity

Show:  this  disjoints  named sub/superclasses

Found 9 uses of NonIdentifiableIdentity

- AnonymousIdentity**
  - AnonymousIdentity **subClassOf** NonIdentifiableIdentity
- IdentifiableIdentity**
  - IdentifiableIdentity **disjointWith** NonIdentifiableIdentity
- Identity**
  - Identity **equivalentTo** IdentifiableIdentity **or** NonIdentifiableIdentity
- Noidentity**
  - Noidentity **subClassOf** NonIdentifiableIdentity
- NonIdentifiableIdentity**
  - NonIdentifiableIdentity **subClassOf** AnonymousIdentity **or** Noidentity **or** PseudoanonymousIdentity
  - Class(NonIdentifiableIdentity)
  - NonIdentifiableIdentity **subClassOf** Identity
  - IdentifiableIdentity **disjointWith** NonIdentifiableIdentity
- PseudoanonymousIdentity**
  - PseudoanonymousIdentity **subClassOf** NonIdentifiableIdentity

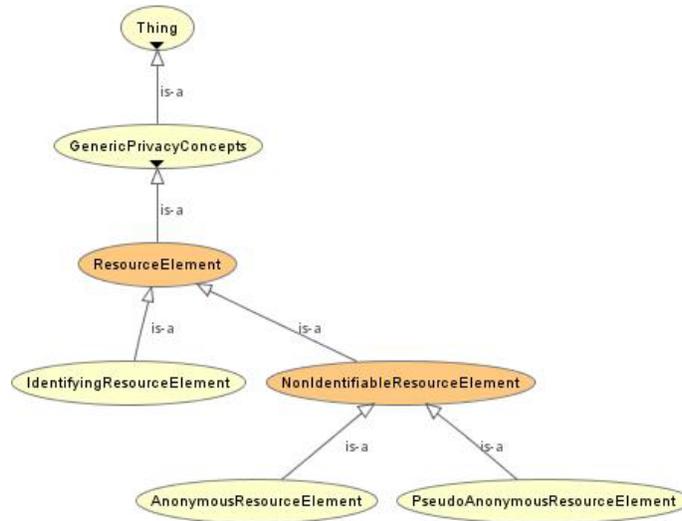
Figure 42: Example: Identities usage

Our next area of concepts is the hierarchy of identities as depicted in Figure 41 and Figure 42. It shows the kinds of identities we have defined and their usage in the form of pseudo code snippets followed by the object property (relationship) "identifies", which states that a particular resource element can identify a particular type of identity as depicted in Figure 43.

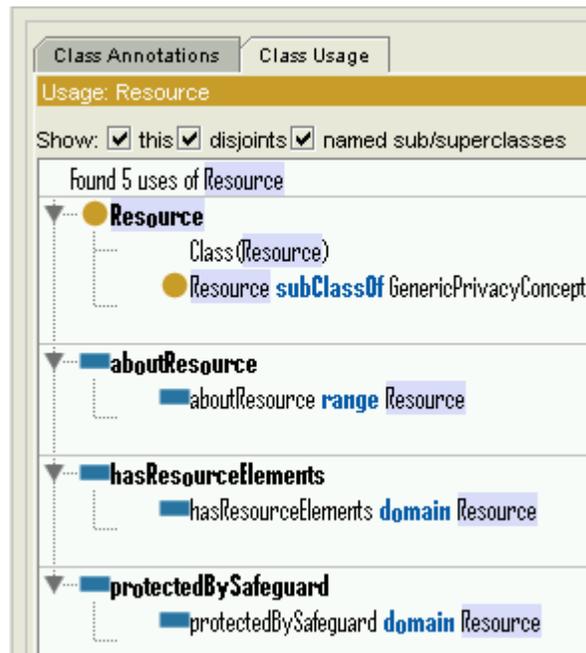


**Figure 43: "identifies" object property**

A further main concept is that of "Resource", which has a number of "ResourceElements" with varying identification characteristics. ResourceElements are not specified as subclasses of Resource, as they do not have their own safeguards. A resource is protected by a safeguard and consists of one or more resource elements as shown in Figure 44 and Figure 45.



**Figure 44: ResourceElements hierarchy**



**Figure 45: Resource usage**

Finally, we show the concept of policy for this section of main generic privacy concepts. A policy is a resource and applies to a resource. Therefore, it is also protected by a safeguard. A policy is composed of a number of statements, which are about a resource, have a recipient, a purpose and a retention as shown in Figure 46 in the form of pseudo code.



Figure 46: Statement concept

#### 8.4.2. Privacy process

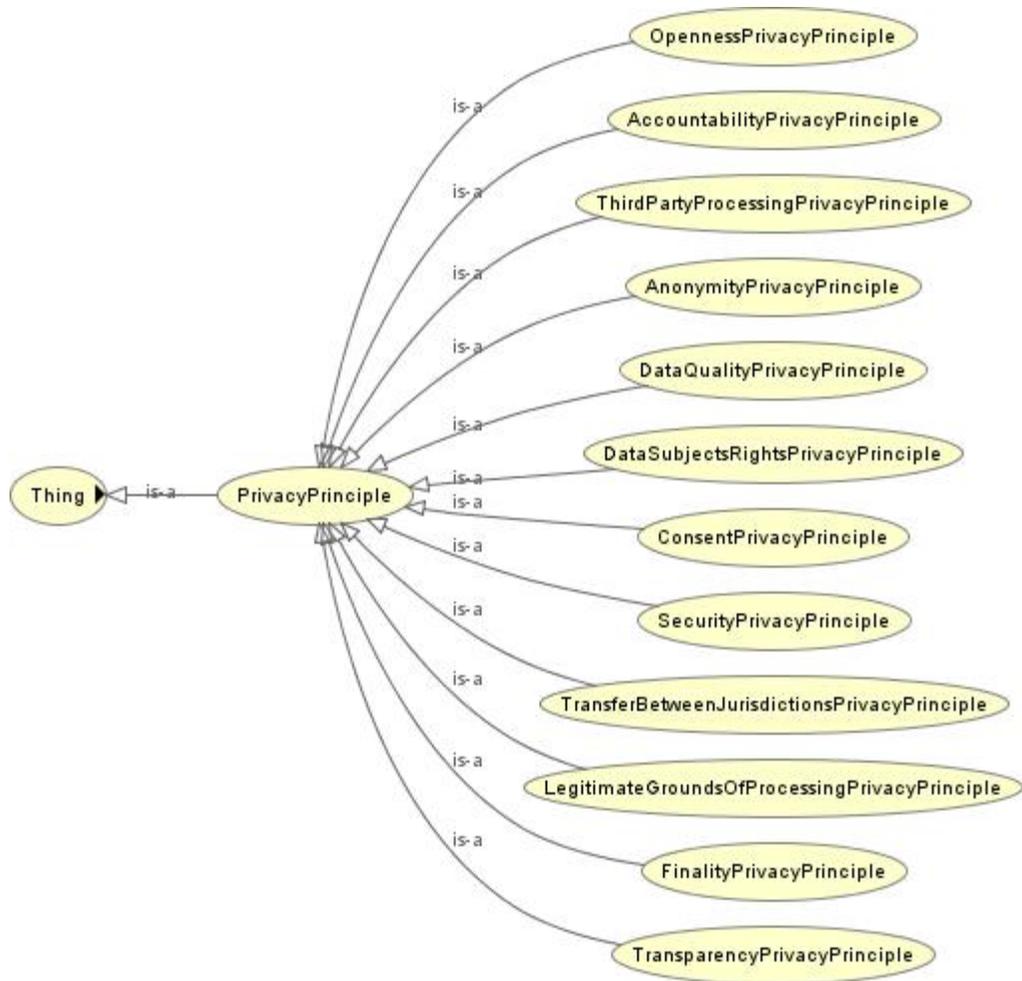
The notion of privacy process has been described in previous chapters and defines an abstract concept of PrivacyProcess with a number of object datatypes (relationships) that define the resources about which this privacy process is all about, a policy that governs this process, an entity that performs this process and a safeguard that protects this process. In general, a "TransitSafeguard" would be applied to privacy processes; however, we cannot exclude the possibility of processes that require other types of safeguards in future extensions and hence, keep it more general as shown in Figure 47.



Figure 47: Privacy Process

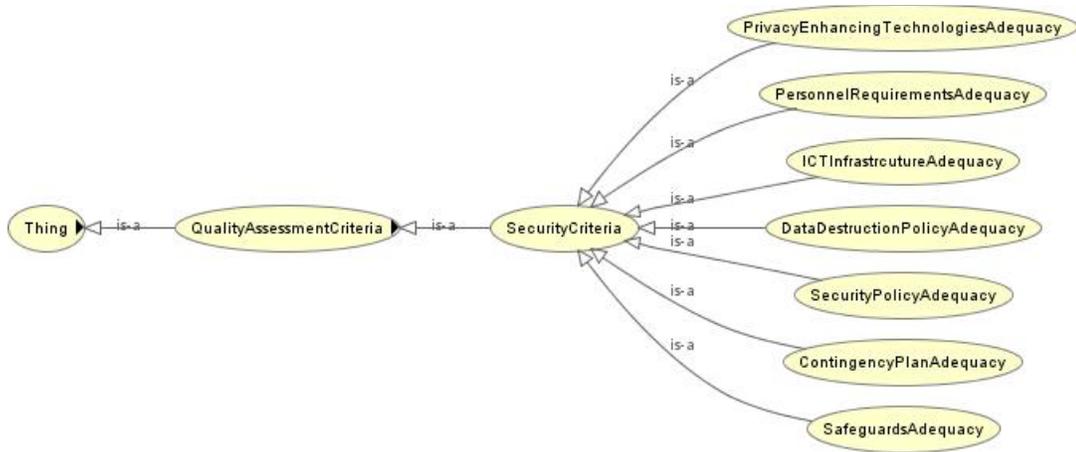
### 8.4.3. Privacy principles

We have defined a number of privacy principles, which are also known as quality aspects and shown in Figure 48. They are used to determine the overall level of privacy one experiences. Every quality aspect has a number of different quality assessment criteria to evaluate the level of privacy for this particular principle.



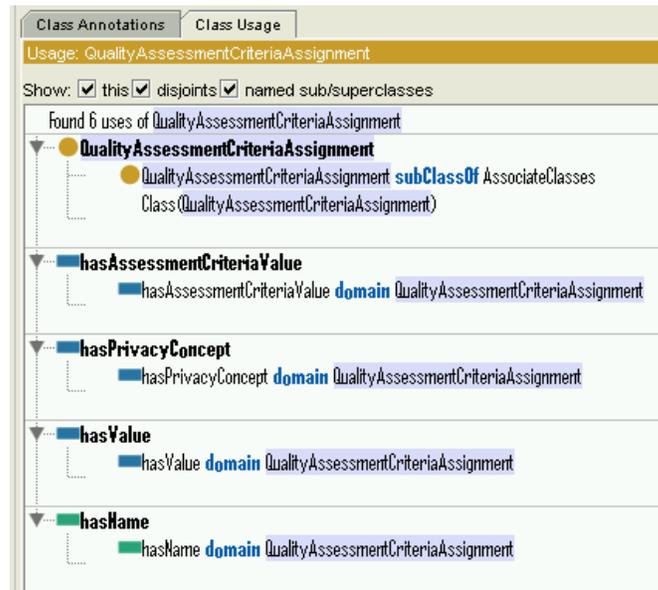
**Figure 48: Privacy principles / quality aspects**

For every privacy principle, we have a pool of classes that represent the quality assessment criteria. An example of these is shown in Figure 49, which depicts the SecurityCriteria that contain the different assessment criteria for the SecurityPrivacyPrinciple.



**Figure 49: SecurityCriteria concepts**

In order to provide a facility that allows the assignment of concepts from our range of quality assessment criteria levels (see next section), we have chosen to use an associate class as this would involve either a triangle relationship or would require OWL to allow for properties of properties, which is not permitted. Hence, we have created an associate class named "QualityAssessmentCriteriaAssignment", which has three object and one datatype property as shown in Figure 50.

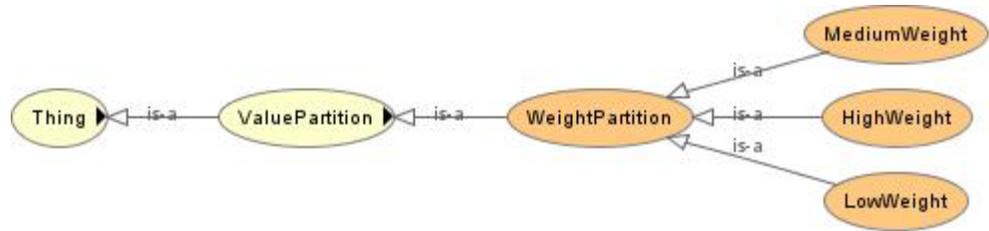


**Figure 50: Associate class**

The first object property is "hasPrivacyConcept", which states that this associate class is linked to some concept (including processes) from the "GenericPrivacyConcepts" domain. The second object property is named "hasAssessmentCriteriaValue", which links it to the domain of concepts named "QualityAssessmentCriteria", which in turn are the concepts that provide the quality assessment criteria for the quality aspects. The third and final object property is named "hasValue", which links it to the domain of concepts named "QualityAssessmentCriteriaLevelRange", which is the partition of concepts that defines the possible assertions for any given quality assessment criteria. The datatype property is a string, which defines an English term defining the actual meaning our assignment. For example, it could have a meaning of "Medium" or any other arbitrary string, if the concept of "MediumLevel" is assigned to the quality assessment criteria "RelevanceToPurpose", subclass of the "DataQualityCriteria" concept.

#### 8.4.4. Quality assessment criteria values

Our privacy evaluation as described in previous chapters is based upon two different evaluation principles. Firstly, we have the domain specific weight, which specifies the "importance" of a certain quality aspect in the domain. This is modeled as a value partition and can have three members, High, Medium and Low, which are mutually disjoint as shown in Figure 51.



**Figure 51: Weight partition**

As described, OWL Viz does not allow us to show the datatype properties that we have assigned to our value partition. Therefore, we show them as a screenshot of the data property usage windows, which is shown in Figure 52 and shows that our concepts of Low, Medium and High have assigned values of 0.1, 0.5 and 0.9 respectively.

Data Property Annotations    Data Property Usage

Usage: hasWeightValue

Show:  this  disjoints

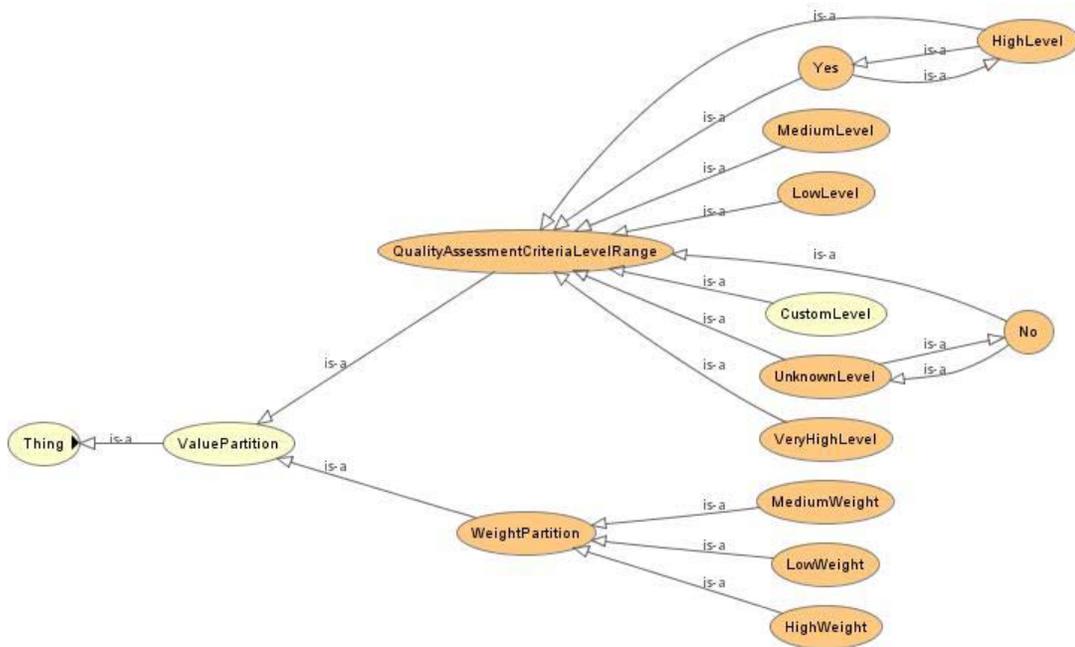
Found 8 uses of hasWeightValue:

- HighWeight**
  - HighWeight **equivalentTo** WeightPartition
  - and hasWeightValue **value** 0.9
- LowWeight**
  - LowWeight **equivalentTo** WeightPartition
  - and hasWeightValue **value** 0.1
- MediumWeight**
  - MediumWeight **equivalentTo** WeightPartition
  - and hasWeightValue **value** 0.5
- WeightPartition**
  - WeightPartition **subclassOf** hasWeightValue **some** double
  - and hasWeightValue **exactly** 1 Literal
- hasWeightValue**
  - Data property (hasWeightValue)
  - hasWeightValue **domain** WeightPartition
  - hasWeightValue **range** double
  - functional:** hasWeightValue

**Figure 52: "hasWeightValue" datatype property**

The second part of the privacy evaluation process to determine the level of privacy involves ascertaining the impact that concepts and their relationships and attributes, as well as processes, have on various

quality assessment criteria of the quality aspects. This is modeled by a value partition as well and is shown in Figure 53, which has a number of members that are mostly mutually exclusive. We also provide some synonyms as some of the concepts are regarded as equal if they have the same value instances. The datatype properties of this value partition are described in Figure 54. It is noteworthy, that synonyms do not appear in that picture, as their usage is inferred automatically, due to the equivalence relationship with other concepts. The values assigned to these assessment criteria are ordinal numbers and range from one to five.



**Figure 53: Quality Assessment Criteria Value Partition**



Figure 54: hasStars datatype property

## 8.5. Conclusion

In this section, we have provided screenshots from our modeling tool Protégé as well as pseudo code snippets in the form of class or property usages, which are provided by Protégé. Our class hierarchy has been organized as shown in Figure 55, which shows the various sub-domains of our conceptualization. This includes the associate class as mentioned above, the quality assessment criteria and their assessed privacy principles

as well as the generic privacy concepts that include the various concepts of the generic privacy ontology. Finally, it contains the value partitions that are used to evaluate ordinal and fuzzy values to our assessment criteria.

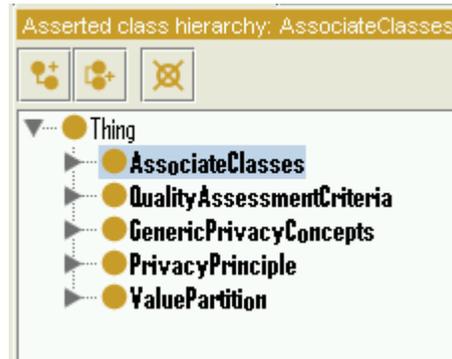


Figure 55: Class hierarchy

## 9. Summary of thesis

### 9.1. Introduction

This chapter is divided into two sections. In the first section, we will give a recapitulation of the work carried out in the thesis. We will carefully examine our research questions posed in Chapter 3 while doing this and discuss how the work contributes to addressing them. This is then followed by a discussion of future work that flows from this thesis.

### 9.2. Recapitulation

In Chapter 1, we introduced the general notion of privacy and discussed how it differs significantly from the concept of security, although privacy has been mistaken for security for too long. We then gave a broad overview of the field and described how privacy differs from individual to individual and from domain to domain. This was continued with a brief elaboration of the problems encountered in the area of privacy and we showed some examples of breaches where personal information got "lost" or was sold or shared without the consent of the subjects affected. This was supported by examining the financial impact that a breach of privacy has for the affected company or individual. Next, we provided a motivational and explanatory discussion of the domain of e-commerce and noted how privacy and the potential loss of privacy can be an inhibiting factor for its growth. Here, we identified some of the key motivations which include the simple and easy access to entering information in large and distributed databases and repositories, while losing most of the control once the information entered.

This was followed by a discussion of the issues of privacy by design, which means that privacy enhancing technologies should not be added on top of

existing systems, but should be built into the design directly, which the literature has discussed for more than a decade. This, however, would require system developers to be experts in that domain as current tools and systems cannot provide appropriate assistance in this matter. Finally, we defined the scope of the problem, which was not to create new or better privacy enhancing technologies such as better encryption tools or other security mechanisms. We defined our scope in a much broader way in that we had to consider all levels and incorporate numerous concepts from different areas in order to provide a better understanding of privacy and a better way to enhance the user experience in the domain of privacy without limiting ourselves to a particular application domain.

In Chapter 2, we discussed the background of privacy and its evolution over time, starting with the first official mention of privacy in the domain of law and justice in the late nineteenth hundreds. Our discussion revealed different definitions of privacy, depending on the context of its usage. This was followed by the different notions of privacy, which provide different meanings and ways in which privacy can be approached and addressed. We continued with a discussion of the ways in which legislation influences our privacy experience and support and how it varies across different countries and regions. This can also be conflicting due to regional, state and federal laws and regulations. Our background focus then narrowed to privacy issues and challenges on the web, which include the inherent open, non-deterministic nature of the web as well as current trends and techniques to support users and protect their privacy needs. This is mainly driven by P3P, the Platform for Privacy Preferences and its related preference exchange language (APPEL). Subsequently, we narrowed our focus even more and looked at the semantic web and the inherent problems and issues in the context of privacy, demonstrating how ontologies provide meaning to the

semantic web to allow for automatic reasoning. We concluded the chapter with a brief look at some of the current research trends in terms of privacy preserving data mining, which provides solutions such as k-Anonymity or l-Diversity that are nevertheless limited to some extent.

In Chapter 3, we stated the problem definition by providing our definitions of concepts that have been used throughout the thesis. They include our own definition and understanding of privacy as well as a discussion of the different notions of privacy, namely the right data, the right purpose and access control. Other definitions included trust and reputation and security and safeguards. This was followed by a definition of the main terms in privacy, namely data subject, entity, resource and process amongst others, and continued with an elaboration of the different dimensions of privacy, which we also refer to as privacy principles. They include data quality, anonymity, transparency, finality principle and security, to name a few. The problem definition followed and explained that privacy is not just a technical problem, but involves legal and jurisdictional dimensions. Hence, we state that we address the problem of capturing the knowledge in the field of privacy, with its concepts and relationships, which allows for semi-automatic processing. This has led us to our research questions, which we re-iterate here:

- 1) "How do we represent the various privacy concepts and relationships and the way they link up with each other?"
- 2) "How can such a representation be used to increase the level of privacy preservation for the data subject?"
- 3) "How can such a representation be used by system developers and service providers to achieve compliance with current rules and regulations as well as state of the art techniques to safeguard personal information?"

4) "How can this representation be used throughout various application domains and provide cross-domain privacy preservation experiences?"

These research questions prompted us to consider the different levels and to determine privacy preservation requirements and the level of privacy experienced by individuals and its necessity to represent the knowledge in the domain within a conceptual framework. Before such a conceptualization could be commenced, it was necessary to discuss the methodology chosen to solve the problem. We looked at Natural Language Processing (NLP), which is unsuitable due to the difficulty of detecting the hidden knowledge in plain and unstructured text. We also looked at a mathematical representation, which was also deemed unsuitable due to the endogenous nature of privacy, which is uncertain and would require alterations every time new technologies or application domains emerge or evolved. We then concentrated on the study of knowledge representation and had the choice of a glossary of terms, which provides no taxonomies among the concepts or any kind of relation between the concepts, which is required for our purpose of capturing and inferring the knowledge in this domain. Topic Maps were another possible methodology and although promising, this was inadequate for our purpose, which was to define the allowable types in order to categorize and evaluate privacy preserving mechanisms. Eventually, we investigated the area of ontologies, which are defined as "a formal shared, explicit, but partial specification of the commonly agreed upon intended meaning of a conceptualisation", among many other definitions. This approach seemed appropriate for our requirements of representing knowledge in a particular domain and structuring it appropriately and hence, addressing our research question 1) with a satisfactory elucidation. In order to build an ontology however, a viable

methodology for its creation had to be found and the METHONTOLOGY approach was chosen for our work.

In Chapter 4, we provided a brief solution overview by describing the various components involved, which can be seen as a guide to understand them and their interactions. We re-iterated the idea of an ontology to address our problems and depicted the real world by showing that the real world has numerous application domains, which all have a privacy component. Furthermore, privacy is its own domain that intersects with all application domains to some extent. Therefore, we concluded that it must be possible to model the application domain unspecific concepts and ideas into a separate concept domain. We started to model two different ontology concepts: firstly, a core ontology that represents the core concepts of the domain of privacy and secondly, specialized ontologies which represent commitments to the concepts of the generic privacy ontology but in addition describe specialized constraints, concepts and relationships pertinent to the application domain.

Having developed the idea of such a generic and specialized ontology, it was necessary to find a means of measuring the level of privacy a user experiences. This need was addressed with the concepts of privacy principles which are quality aspects or dimensions to (measure) privacy. Hence, we concluded that the various concepts in our generic privacy ontology would determine the quality aspect that would be affected if that concept were to be used. However, at this stage, it would not be possible to determine the actual level of that influence as this is an application domain specific property. Therefore, application specific extensions of the generic privacy ontology were required. These extensions have to be created by an expert of the application domain, who uses concepts from that domain and

"re-attaches" them to the generic privacy ontology, creating an application domain privacy ontology that commits to all the concepts and requirements of the generic one and to a subset of the application domain ones, as only the ones that are pertinent to privacy need to be included. This kind of generic and application specific ontology makes it possible to provide support for privacy on inter-application domain specific levels as all the concepts, attributes and relationships of the application domain privacy ontology are inferred back to their counterparts of the generic privacy ontology. This essentially addresses our research question 4.

Chapter 4 concluded with an elaboration of the main concepts of the generic privacy ontology, namely an entity, the data subject and a resource, which is a concept that represents personal information about a data subject.

In Chapter 5, provided the concepts, properties and relationships of the generic privacy ontology at a far greater level of detail. We started with the hierarchy of entities, which is essentially a hierarchy defining different levels of access to resources about data subjects. These resources could also be categorized further into resources that have identifying characteristics and others that have pseudo identifying ones or none at all. However, as entities and hence, data subjects, might not want to be identified directly in every instance, identities were introduced that could represent the entity altogether, to an extent or not at all and it is the entity's choice to determine which identity to use. This includes the choice of remaining anonymous in certain transactions if necessary. It was also clear that resources need to be protected by safeguards, which can be classified into safeguards for transit or non-transit needs as well as access control safeguards. Furthermore, the ontology required support for policies, such as

privacy policies to support the data subjects with their decision making in terms of access control to their personal information. We have shown some of the common purposes such a privacy policy can have, adapting from P3P. In many cases, personal information is involved in processes, for example, when entering them into a system or when they are processed by an entity. This led to the need for the conceptual model of a privacy process, which is a process that is about personal information, governed by a policy, protected by a safeguard and performed by an entity. An example of such a privacy process is the shared resource process, which is a sub-concept of the privacy process and adds properties like recipient and sharedBy.

This chapter continues with an elaboration of the representation of the privacy principles and their quality assessment criteria, which are used to determine the level of privacy for a particular transaction or process. As privacy is an endogenous concept, mathematically precise values were not appropriate and therefore, we chose to use ordinal values from a range of one to five, where one is the lowest and five the highest value, which is represented by one to five stars. We have determined the actual level assignments for each of the quality assessment criteria and assigned possible values to them, which are used to determine the level of privacy. After this elaboration, we have provided the formulas to determine the level of privacy in a given context. However, as the generic privacy ontology is less concrete, a calculation can be made only in the application domain specific extension. As our basic privacy concepts are the same across all application domains, but their needs for protection may be different, we introduced a weight factor, which is chosen initially by the domain expert to specify the impact factor of a particular privacy principle in an application domain specific extension. For example, unlike the financial domain, in the medical domain, safeguards are treated in a more circumscribed manner

and allow for exceptions, as access to personal information in a medical emergency must still be possible. Therefore, choosing a softer safeguard in the medical domain, but with a lower impact factor of the associated privacy principle, leads to a smaller output factor when determining the level of privacy.

To make our idea more specific, we chose two different application domain examples, in Chapter 6 and 7, describing the medical and the e-commerce domains respectively. In both these example chapters, we begin by discussing the concepts of the respective application domains and describe examples of possible processes in that domain. However, due to the complexity and sheer size of any of these domains, we had to limit ourselves to restricted sub-domains of these application domains.

In the example of the restricted medical domain given in Chapter 6, we discuss a possible hospital scenario. The prime concepts in such a scenario are medical provider, patient, staff and patient, record amongst others. Our example concept of "process" in this domain is an admission process where a patient is admitted to a medical provider for a particular reason. Upon this base, we modeled the medical privacy ontology by using the concept from our restricted medical domain and in conjunction with the concepts of the generic privacy ontology, creating a specialization of the latter. For example, a patient was classified as "AliveDataSubject\_Patient" in our specialization. After describing the model of our application domain specific privacy ontology, we demonstrated some of the instances and hence, the various influences these concepts and instances have on the privacy principles. Additionally, we provided the weight factors of the medical domain in tabular form and used it together with our influence levels to determine the level of privacy for our hospital scenario.

Our approach in Chapter 7 was similar to the one in Chapter 6; however, we concentrated on the designer or developer perspective instead of the user perspective. The e-commerce domain example was used to incorporate privacy concepts into the design of a new e-commerce shopping system while it is being built in order to allow the engineers of that system to adhere to industry standards and legislation with regards to privacy. As engineers and system developers are not necessarily experts on the issue of privacy, they use this ontology to gain a greater understanding of its concepts and dimensions as well as their influences on the overall level of privacy that the customers experience. This is beneficial for both customers and service providers as both can be confident that personal information is protected appropriately and financial and image loss is less likely to occur from loss or misuse of personal information.

After having described the various concepts, attributes and relationships of the application domain, a restricted B2C e-commerce one, we selected a few processes that the system had to be able to support. They included a user registration process, an order and payment, and a delivery process. We then created our application domain specific specialization of the generic privacy ontology, analogous to the approach used in Chapter 6. However, instead of simply evaluating the level of privacy for given concepts, we examined concepts that would support and enhance the level of privacy within our development. Therefore, we selected mainly those concepts that would highly and positively influence the quality assessment criteria and their related quality aspects. Eventually, a system emerged that included mainly concepts that would support and strengthen the level of privacy for the user and, hence, helped us to address research question 3).

In Chapter 8, we discussed our choice of language for our technical implementation. We discussed RDF and RDF schema as well as the different flavours of OWL, name Lite, DL and Full and noted that DL is the only viable option due to the limitations of Lite and the lack of computational guarantees for OWL-Full. In our elaboration, we showed many of the concepts and properties of our generic privacy ontology and the privacy principles as they have been implemented with the help of our implementation tool Protégé.

### 9.3. Future Work

#### 9.3.1. Inclusion of Legal Frameworks

Throughout this work, due to time limitations, we carefully circumscribed the problem to be addressed within the scope of this thesis. One issue that should be included in future work is the support for a legal framework. In order to fully understand the implications of a territory and its jurisdiction for the protection of user experienced privacy, it would be necessary to model them in a conceptual and semantic framework to allow for automatic reasoning within our domain. In general, it might be possible to model these on an ontological basis and integrate them within the generic privacy ontology. This kind of integration needs to be researched further, also with regards to other ontologies, such as those relating to trust and reputation. The integration or a mapping of these would greatly enhance the privacy effect and the overall experience that the data subject will encounter. Furthermore, more work needs to be done to develop a mapping to upper ontologies to provide more support for high level abstract terms and, hence, the integration with other systems.

### 9.3.2. Automating through an MDA architecture

Although our work here to support developers and engineers of systems to gain a greater understanding of privacy and build the relevant principles into their systems was fruitful, it still contains many manual steps as described in Chapter 7. Our future research direction is a model-driven architecture, which is guided by the generic privacy ontology and its extensions to application domains, to automate those processes. We believe that it is possible to create a visual designer or editor that can be used to create programs and enrich them with privacy concepts on a fundamental level. This may be an extension or plugin to Eclipse, but as yet is unknown.

### 9.3.3. Integration with Security Ontology

In our discussion, we have mentioned the work of Schumacher [39], who proposed a security ontology to model the domain concepts of security. As elaborated, security and in particular, safeguards, are a vital component of privacy protection. Hence, the privacy ontology would benefit from an inclusion of such a security ontology to model safeguards and their associated properties effectively.

### 9.3.4. Integration with Privacy Preserving Databases

Our generic privacy ontology and its extensions to specific application domains could be used and integrated with privacy preserving databases in multiple ways. the ontology has concepts that represent privacy preserving databases, as these are one of the safeguards to preserving personal information. Different types of privacy preserving

databases are represented by their respective concepts and in turn linked to the different privacy principles to determine how well such a database can preserve privacy and therefore the privacy level experienced can be determined. Secondly, the ontology could be used by the database management system directly in order to determine the level of privacy desired for particular sets of information. This would require every piece of information or groups of information in the database to be linked to their ontology concept counterpart in order to classify the data and extract the rules and policies associated with it. This work of integrating the ontology into the database management system is beyond the scope of this thesis and hence, subject of future work.

#### 9.3.5. Monotonic Process Changes

A useful theoretical element that can be added to the current work is characterizing monotonic process changes with regards to the level of privacy. It entails the issue of keeping the level of privacy at the same level while making changes to elements of processes that have been developed. Therefore, it is necessary to look at the various elements that will form part of the modified process. This will be addressed in future work as it is beyond the scope of this thesis and will be researched and investigated at much greater level of detail.

#### 9.3.6. Correlation of Privacy Rules with the Privacy Ontology

In real world as well as digital world scenarios, privacy rules have been established that are normally composed of natural language, but may also be composed of languages such as P3P. In order for them to be

usable or re-usable within the privacy ontology, the rules would need to be conceptualized as concepts and relationships of the ontology itself and structured as statements in the policy part of the ontology. Once this is done, the rules have semantic value and can be used within the ontology to aid in privacy preservation. An implementation of such a conversion from rules to statement concepts is subject to future work as well as the relevant evaluation of this conversion with regards to the complexity, expressiveness and semantic clarity.

#### 9.4. Conclusion

We conclude this thesis by briefly recapitulating our main goals and achievements. We have created a generic privacy ontology as well as extensions to it and the ability to extend the generic one to cater for other application domains. These can be used to determine the level of privacy for certain concepts, and processes within the application domain by the data subject, application designers or developers, or other entities that are involved with processing personal information in any given situation where the ontology is used. In future work, we will be demonstrating scenarios in a more comparative fashion to elaborate the value of the privacy ontology, describing how they would work out with and without support of the ontology. This will show even better the different levels of privacy experienced when using and sharing personal information.

## 10. References

1. IdentityTheftResourceCenter, "2008 Data Breach Totals Soar," 2009; [http://www.idtheftcenter.org/artman2/publish/m\\_press/2008\\_Data\\_Breach\\_Totals\\_Soar\\_printer.shtml](http://www.idtheftcenter.org/artman2/publish/m_press/2008_Data_Breach_Totals_Soar_printer.shtml).
2. A.I. Anton, et al., "Inside JetBlue's privacy policy violations," *Security & Privacy Magazine, IEEE*, vol. 2, no. 6, 2004, pp. 12-18.
3. R. Singel, "JetBlue Shared Passenger Data," 2003; <http://www.wired.com/politics/security/news/2003/09/60489>.
4. Consumers\_Union, "Another week, another identity theft scandal: Recent Data Security Breaches Underscore Need for Stronger Identity Theft Protections," 2005; <http://www.consumersunion.org/creditmatters/creditmattersupdates/002244.html>.
5. Privacy\_Rights\_Clearinghouse, "A Chronology of Data Breaches," 2009; <http://www.privacyrights.org/ar/ChronDataBreaches.htm>.
6. W. Grayson, "UA says probe continues of '08 hacking," 2009; [http://www.tuscaloosaneews.com/article/20090214/NEWS/902130209/1007?Title=UA\\_says\\_probe\\_continues\\_of\\_\\_08\\_hacking](http://www.tuscaloosaneews.com/article/20090214/NEWS/902130209/1007?Title=UA_says_probe_continues_of__08_hacking).
7. Information\_Shield, "Privacy Breach Impact Calculator," 2007; <http://www.informationshield.com/privacybreachcalc.html>.
8. A. Cavoukian, *Privacy By Design*, Information and Privacy Commissioner of Canada, 2009.
9. S.D. Warren and L.D. Brandeis, "The Right To Privacy," *Harvard law review*, vol. 4, 1890, pp. 193-220.
10. A.F. Westin, *Privacy and freedom*, The Bodley Head Ltd, 1970.
11. E. Chang, et al., *Trust and Reputation for Service-Oriented Environments*, John Wiley & Sons Ltd, 2006.
12. Privacilla, "Privacy Fundamentals: Privacilla's Two-Part Definition of Privacy," 2003; <http://www.privacilla.org/fundamentals/privacydefinition.html>.
13. G. Radwanski, "Privacy," *Book Privacy*, Series Privacy, ed., Editor ed.^eds., 2002, pp.

14. D. Banisar, "Data Protection Laws Around The World," 2004;  
<http://www.privacyinternational.org/survey/dpmap.jpg>.
15. R.F.C. Bouchard and J.D.C. Franklin, *Guidebook to the Freedom of Information and Privacy Acts [and] 1981 Supplement*, Clark Boardman Company, Ltd., 435 Hudson Street, New York, NY 10014., 1980.
16. , "Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data," *Official Journal of the European Communities (OJEC)*, vol. L. 281, pp. 31-50.
17. Federal\_Communications\_Commission, "The Wireless Communications and Public Safety Act of 1999 (911 Act)," 1999.
18. G. Karjoth, et al., "A privacy policy model for enterprises  
A privacy policy model for enterprises," *Proc. Computer Security Foundations Workshop, 2002. Proceedings. 15th IEEE*, 2002, pp. 271-281.
19. A.R.A. Bouguettaya and M.Y. Eltoweissy, "Privacy on the Web: facts, challenges, and solutions," *Security & Privacy Magazine, IEEE*, vol. 1, no. 6, 2003, pp. 40-49.
20. PriceWaterhouseCoopers, "E-Privacy: Solving the On-Line Equation," 2002;  
[www.pwcglobal.com/extweb/pwcpublications.nsf/](http://www.pwcglobal.com/extweb/pwcpublications.nsf/).
21. Wikipedia, "Federated Identity," 2007;  
[http://en.wikipedia.org/wiki/Federated\\_identity](http://en.wikipedia.org/wiki/Federated_identity).
22. L. Cranor, et al., *The Platform for Privacy Preferences 1.0 (P3P1.0) Specification*, World Wide Web Consortium, 2002.
23. L. Cranor, et al., *A P3P Preference Exchange Language 1.0 (APPEL1.0)*, World Wide Web Consortium, 2002.
24. R. Agrawal, et al., "XPref: a preference language for P3P," *Computer Networks*, vol. 48, no. 5, 2005, pp. 809-827.
25. M. Hecker, "A privacy management system for mobile devices in context-aware environments," Faculty of Information Technology, University of Technology, Sydney, Australia, 2004.

26. T.R. Gruber, "A translation approach to portable ontologies," *Knowledge Acquisition*, vol. 5, no. 2, 1993, pp. 199-220.
27. N. Guarino and P. Giaretta, "Ontologies and Knowledge Bases: Towards a Terminological Clarification," *Towards Very Large Knowledge Bases: Knowledge Building and Knowledge Sharing* 1995.
28. P. Spyns and R. Meersman, "Ontology engineering and (digital) business ecosystems: A case for a pragmatic web," *Proc. Emerging Technologies and Factory Automation, 2007. ETFA. IEEE Conference on*, 2007, pp. 831-838.
29. A. Kim, et al., "Building Privacy into the Semantic Web: An Ontology Needed Now," *Book Building Privacy into the Semantic Web: An Ontology Needed Now*, Series Building Privacy into the Semantic Web: An Ontology Needed Now, ed., Editor ed.^eds., 2002, pp.
30. B. Thuraisingham, "Confidentiality, Privacy and Trust Policy Enforcement for the Semantic Web," *Proc. Policies for Distributed Systems and Networks, 2007. POLICY '07. Eighth IEEE International Workshop on*, 2007, pp. 8-11.
31. H. Alani, et al., "Building a Pragmatic Semantic Web," *Intelligent Systems, IEEE*, vol. 23, no. 3, 2008, pp. 61-68.
32. M. Ashburner, et al., "Creating the Gene Ontology Resource: Design and Implementation," *Genome Research*, vol. 11, 2001, pp. 1425-1433.
33. A.S. Sidhu, "Structured Vocabularies for Proteins: Protein Ontology Project," Faculty of Science, Technology and Engineering, La Trobe University, Melbourne, 2008.
34. G. Denker, et al., "Security for DAML Web Services: Annotation and Matchmaking," *The Semantic Web - ISWC 2003*, 2003, pp. 335-350.
35. A. Tumer, et al., "A Semantic based Privacy Framework for Web Services," *Book A Semantic based Privacy Framework for Web Services*, Series A Semantic based Privacy Framework for Web Services, ed., Editor ed.^eds., 2003, pp.
36. L. Qin and V. Atluri, "Concept-level access control for the Semantic Web," *Book Concept-level access control for the Semantic Web*, Series Concept-level access control for the Semantic Web, ed., Editor ed.^eds., ACM, 2003, pp.

37. J. Castro, et al., "A Requirements-Driven Development Methodology," *Advanced Information Systems Engineering*, 2001, pp. 108-123.
38. H. Mouratidis, et al., "An Ontology for Modelling Security: The Tropos Approach," *Knowledge-Based Intelligent Information and Engineering Systems*, 2003, pp. 1387-1394.
39. M. Schumacher, "Security Engineering with Patterns: 6. Toward a Security Core Ontology," *Lecture Notes in Computer Science* 2754, 2003, pp. 87-96.
40. E. Gamma, *Design patterns : elements of reusable object-oriented software*, Addison-Wesley, 1995, p. xv, 395 p.
41. Y. Sure and J. Haller, "Towards Cross-Domain Security Properties Supported by Ontologies," *Web Information Systems – WISE 2004 Workshops*, 2004, pp. 58-69.
42. L. Kagal, et al., "Authorization and privacy for semantic Web services," *Intelligent Systems, IEEE*, vol. 19, no. 4, 2004, pp. 50-56.
43. L. Kagal, et al., "A Policy Based Approach to Security for the Semantic Web," *The SemanticWeb - ISWC 2003*, 2003, pp. 402-418.
44. R. Agrawal and S. Ramakrishnan, "Privacy-preserving data mining," *SIGMOD Rec.*, vol. 29, no. 2, 2000, pp. 439-450; DOI <http://doi.acm.org/10.1145/335191.335438>.
45. D. Agrawal and C.C. Aggarwal, "On the design and quantification of privacy preserving data mining algorithms," *Proc. Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, ACM, 2001.
46. S.I. Vijay, "Transforming data to satisfy privacy constraints," *Book Transforming data to satisfy privacy constraints*, Series Transforming data to satisfy privacy constraints, ed., Editor ed.^eds., ACM, 2002, pp.
47. R.J. Bayardo and A. Rakesh, "Data privacy through optimal k-anonymization," *Proc. Data Engineering, 2005. ICDE 2005. Proceedings. 21st International Conference on*, 2005, pp. 217-228.
48. C.C. Aggarwal, "On *k*-anonymity and the curse of dimensionality," *Book On *k*-anonymity and the curse of dimensionality*, Series On *k*-

anonymity and the curse of dimensionality, ed., Editor ed.^eds., VLDB Endowment, 2005, pp.

49. L. Kristen, et al., "Incognito: efficient full-domain K-anonymity," *Book Incognito: efficient full-domain K-anonymity*, Series Incognito: efficient full-domain K-anonymity, ed., Editor ed.^eds., ACM, 2005, pp.

50. Z. Sheng, et al., "Privacy-enhancing *k*-anonymization of customer data," *Proc. Proceedings of the twenty-fourth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, ACM, 2005.

51. M.E. Nergiz and C. Clifton, "Thoughts on k-Anonymization," *Proc. Data Engineering Workshops, 2006. Proceedings. 22nd International Conference on*, 2006, pp. 96-96.

52. X. Xiaokui and T. Yufei, "Anatomy: simple and effective privacy preservation," *Proc. Proceedings of the 32nd international conference on Very large data bases*, VLDB Endowment, 2006.

53. F. Arik, et al., "Providing k-anonymity in data mining," *The VLDB Journal*, vol. 17, no. 4, 2008, pp. 789-804; DOI <http://dx.doi.org/10.1007/s00778-006-0039-5>.

54. A. Machanavajjhala, et al., "l-Diversity: Privacy Beyond k-Anonymity," *Proc. Data Engineering, 2006. ICDE '06. Proceedings of the 22nd International Conference on*, 2006, pp. 24-24.

55. C.C. Aggarwal, "On Randomization, Public Information and the Curse of Dimensionality," *Proc. Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on*, 2007, pp. 136-145.

56. C. Dwork, "Differential Privacy," *Automata, Languages and Programming*, 2006, pp. 1-12.

57. C. Dwork, "Differential Privacy: A Survey of Results," *Theory and Applications of Models of Computation*, 2008, pp. 1-19.

58. Webster\_Dictionary, "Privacy."

59. P. Sieghart, *Privacy and computers*, Latimer New Dimensions, 1976.

60. J.B. Young, *Privacy*, Wiley, 1978.

61. R. Leenes, et al., *PRIME White Paper v2*, white paper, Privacy and Identity Management for Europe, 2007.
62. Wikipedia, "Topic Maps," [http://en.wikipedia.org/wiki/Topic\\_Maps](http://en.wikipedia.org/wiki/Topic_Maps).
63. A. Gómez-Pérez, et al., *Ontological engineering : with examples from the areas of knowledge management, e-commerce and the semantic Web*, Springer-Verlag, 2004, p. xii, 403 p.
64. , *OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data*, Directorate for Science, Technology and Industry, [http://www.oecd.org/document/18/0,2340,en\\_2649\\_34255\\_1815186\\_1\\_1\\_1\\_1,00.html](http://www.oecd.org/document/18/0,2340,en_2649_34255_1815186_1_1_1_1,00.html).
65. J. Huizenga, *Handbook of Privacy and Privacy-Enhancing Technologies: The case of Intelligent Software Agents*, College bescherming persoonsgegevens, 2003.
66. M. Hecker and T.S. Dillon, "Towards a Privacy Ontology," 2005.
67. P. Spyns, et al., "Data modelling versus ontology engineering," *SIGMOND*, vol. 31, no. 4, 2002, pp. 12-17.
68. C. Wouters, "A Formalization and Application of Ontology Extraction," Faculty of Science, Technology and Engineering, La Trobe University, Bundoora, Victoria, Australia, 2005.
69. M. Hecker and T.S. Dillon, "Ontological privacy support for the medical domain," *Proc. Proceedings of eHPass National e-Health Privacy and Security Symposium*, Queensland University of Technology, Brisbane, Australia, 2006.
70. M. Hecker, et al., "Privacy Ontology Support for E-Commerce," *IEEE Internet Computing*, vol. 12, no. 2, 2008, pp. 54-61.
71. H. Chan, et al., *E-Commerce: Fundamentals and Applications*, John Wiley & Sons Ltd, 2001.
72. F. Manola and E. Miller, "RDF Primer," 2004; <http://www.w3.org/TR/2004/REC-rdf-primer-20040210/>.
73. D.L. McGuinness and F.v. Harmelen, "OWL Web Ontology Language Overview," 2004; <http://www.w3.org/TR/owl-features/>.

74. L.A. Stein, et al., “DARPA Agent Markup Language (DAML) Reference,” 2002; <http://www.daml.org/language>.
75. I. Horrocks, et al., “The Ontology Interference Layer (OIL),” 2000; <http://www.ontoknowledge.org/oil/>.
76. I. Horrocks, et al., “DAML+OIL,” 2001; <http://www.daml.org/2001/03/daml+oil-index>.
77. M. Horridge, “A Practical Guide To Building OWL Ontologies With The Protege-OWL Plugin,” 2004.

*Every reasonable effort has been made to acknowledge the owners of copyrighted material. I would be pleased to hear from any copyright owner who has been omitted or incorrectly acknowledged.*