Universidad Politécnica
de Madrid

Escuela Técnica Superior de Ingenieros Informáticos

Master's degree in Data Science

Master´s Final Project

Data Model for Legal and Ethical Compliance by Design in Data Science Projects

`https://github.com/philippScomparin/LSEID`

Author: Philipp Scomparin
Tutor: Rana Saniei and Víctor Rodríguez-Doncel

Madrid, July 2021

This Master's Final Project has been deposited in the ETSI Informáticos of the Polytechnic University of Madrid for its defense.

Master's final project
Master's degree in Data Science

Title: Data Model for Legal and Ethical Compliance by Design in Data Science Projects

July 2021

Author: Philipp Scomparin
Tutor: Rana Saniei and Víctor Rodríguez-Doncel
Departamento de Inteligencia Artificial
ETSI Informáticos
Universidad Politécnica de Madrid

Acknowledgements

First and foremost, I would like to thank my two supervisors: Víctor Rodríguez-Doncel and Rana Saniei. Without their guidance, vision, encouragement and invaluable suggestions, it would not have been possible to carry out this project.

I would also like to thank my colleagues at the university with whom I have exchanged valuable knowledge and opinions in the last months.

Finally, I thank my friends who always supported me and helped me to keep pushing. Their extraordinary understanding made it a lot easier for me to complete the project.

Abstract

Data is the most important ingredient in any data science project. It is also a fundamental aspect of any software development project. Considering the required data specifications the assessment of legal compliance is critical. It is also necessary to evaluate the social and ethical impact of data to avoid undesirable results and consequences.
This project aims to provide sufficient vocabulary in the form of an ontology that allows data publishers to describe their data in terms of legal, social, and ethical impact. Data publishers should be able to explain the main purpose and the legal basis for collecting the data, the risks associated with data processing, and the ethical assessments that have been performed. With all this information, data scientists, software developers and other interested parties can decide whether the data is suitable for their projects.
The main approach to create the ontology was to follow the Linked Open Terms Methodology. This ontology engineering methodology divides the process into four parts: ontology requirements specification, ontology implementation, ontology publication, and ontology maintenance.
After completing all the necessary steps, an ontology has been created and published on `https://protect.oeg.fi.upm.es/def/lseid`. With this vocabulary data publishers are able to provide all the relevant information about legal, social, and ethical aspects of the data.
This thesis describes in detail the whole process of creating the ontology. It also explains design choices and presents the final result.

Contents

List of Figures

Chapter 1

Introduction

1.1 Motivation

2.5 quintillion bytes of data is currently produced every day[1]. The legal, social and ethical impact of data has been an important topic for a long period of time. However, in the last ten years, the discussion has become louder. One of the reasons for this outcry is the scandal of Cambridge Analytica[2] that has drawn the public attention to the impact of data and its risks.

Nowadays, many companies use data for descriptive, prescriptive and predictive analysis. Data is gathered and used in machine learning algorithms to help them make decisions. This causes not only legal but also social and ethical issues. Regarding the legal aspect it is important that only data is used that has been collected with the consent of the data subjects. Furthermore, the data must be processed and controlled as required by law. On the other hand, social and ethical issues can arise, for example, when the used data is biased. As stated in the article published by The official portal for European data "Open data and data bias", projects that use biased data can lead to discrimination. For example, if a recognition system is trained with data that over-represents Western faces it will fail to recognize people with other backgrounds (i.e. Asian, African, etc.)[3].

To counter the problematic usage of data, the EU has imposed several restrictions, with General Data Protection Regulation (GDPR)[1] being the most important one of the recent years. To ensure that companies comply with the GDPR, the EU sets harsh fines in case of violation. The fines can go up to 20 million Euros or 4% of the annual global turnover, whichever is higher. In addition, an ethical assessment is now required for every EU funded project[4]. To facilitate the assessments of data, the EU has published various guidelines over the last years. There has also been a lot of attention on the impact of AI, and this is where the concept of trustworthy AI comes from. The idea of trustworthy AI is to answer a set of questions to see the impact of an AI project on individuals, organizations, society and environment and accordingly, evaluate the trustworthiness of the AI solution. Other tools to assess the ethical impact of a dataset are, for example, ADAPT Ethics Canvas[2], Data Protection

[1] https://gdpr-info.eu
[2] https://ethicscanvas.org

Impact Assessment (DPIA), and Datasheet for Datasets[3]. All of them will be presented in later chapters of this work.
Publishing data has gone through an evolution in the last years. It started from plain txt or csv files. However, with the promising rise in recent years of linked data, it would be convenient to create ontologies and vocabularies to describe datasets in terms of legal, social and ethical impact based on the previously mentioned tools and guidelines. While some of the legal aspects can already be found in the W3C Data Catalog Vocabulary (DCAT)[4], the ethical aspects are still not present in these types of vocabularies. Vocabularies like DCAT provide the necessary terms to describe datasets. However, they neither include social nor ethical impact.

Public institutions in the EU are mandated to create and maintain data portals. Data Portals are publicly available portals that provide datasets. For example, the data portal of Madrid[5] contains useful data about the city and its administration. The datasets on these portals have metadata associated with them. Often, the metadata is provided by making use of DCAT or the DCAT Application Profile for Data Portals in Europe. Extending DCAT and adding vocabularies to describe legal, social and ethical impact can lead to complete metadata on the data portals and therefore further improve the transparency of the public institutions.

Since social and ethical aspects are becoming more important, a dataset description that excludes them cannot be considered as a complete one. Having a full description of a dataset in linked data that includes legal, social and ethical impact would not only be beneficial to the data creators to show that they have thought of the impact of their data, but it would also be convenient for data consumers to get all the information beforehand in an easy way. Data consumers would be able to see all the risks and recommended as well as discouraged use cases for processing the data. This would allow them to better decide which dataset they should go for. Currently, there is no ontology or vocabulary that addresses these topics specifically. So, creating them would facilitate the process for data creators to include the legal, social and ethical assessment in the dataset description. Furthermore, the GDPR strictly regulates data transmission between different stakeholders, such as joint controllers, data processors or other data recipients. A common set of vocabularies that is understandable by all the involved parties which describes datasets in terms of their legal, ethical and social impact would facilitate the whole data transmission process.

1.2 Objectives

1.2.1 Main Objective

The main objective of this work is to develop the vocabularies and ontologies necessary to describe a dataset in terms of the social, legal and ethical impact that applications exploiting it may cause.

1.2.2 Sub-Objectives

To reach the final goal the following sub-objectives need to be achieved:

[3] https://arxiv.org/pdf/1803.09010v6.pdf
[4] https://www.w3.org/TR/vocab-dcat-2
[5] https://datos.madrid.es/portal/site/egob

1. Investigate the necessary information for assessing the legal impacts of a dataset (considering regulations such as GDPR) and create a vocabulary for it
2. Investigate the necessary information for assessing both, the social impact and the ethical impact of a dataset, and create a vocabulary for them
3. Extend the W3C Data Catalog Vocabulary (DCAT) with the new vocabulary

Chapter 2

Research Methodology

2.1 Information Gathering

To gather all the relevant information, it is necessary to carry out a literature review of official documents published by public and non-public institutions as well as investigate the state of the art in similar scientific projects and works.

One key aspect was to obtain a complete understanding on how social and ethical aspects can be assessed. For this reason several assessments and guidelines were read and taken into consideration being the most important ones the DPIA and The Ethics Canvas. While studying the different assessments and guidelines, it was also necessary to understand the different social and ethical risks of data processing. Reading about these topics in news articles and other type of articles and papers available on different platforms such as Google Scholar.[1] was helpful to get a better idea about the issues.

On the other hand, it was also important to understand the legal situation inside the EU with regards to data. For this particular task, it was important to read parts of the official GDPR document as well as other resources that address GDPR and data privacy issues.

2.2 Ontology Engineering

In order to build an ontology efficiently and in a correct way it was necessary to learn about different available ontology engineering techniques. After carefully studying them, it was decided to use the Linked Open Terms methodology which will be explained in chapter 4.4.1.

2.3 Evaluation Methodology

The evaluation of the ontology has been carried out by using the tools *OnToology*[2] and *OOPS!*[3]. Furthermore, it was checked that all the previously defined competency

[1] https://scholar.google.com
[2] https://ontoology.linkeddata.es
[3] http://oops.linkeddata.es

questions of the ontology can be answered.

Chapter 3

Related Work

This section serves to illustrate existing ontologies that are related to the ontology of this project. On the other hand, it will also include guidelines and templates to assess data in terms of social and ethical impact. The reason to include them is that they are a fundamental part of the project that will be represented in the ontology.

3.1 Related Ontologies

3.1.1 Data Catalog Vocabulary

One possibility to describe datasets is provided by the W3C Data Catalog Vocabulary (DCAT). The goal of DCAT is to become an interoperability standard in the European Union. A system using URIs, controlled vocabularies and ontology models like DCAT, provides the possibility to navigate through data catalogues, aggregate metadata and query it in the same way for each data source. So, DCAT can be defined as a model that provides a standard way of describing catalogues. Catalogues contain useful metadata that help to organize and find resources as well as to integrate information from different sources. Usually they are used for describing datasets. However, they can also be used for other resources such as data services.

The three main component classes of DCAT are: catalog, resource, and distribution. Distribution, dataset and data service are subclasses of the resource class. A catalog contains a pointer, either to a dataset or to a data service, and metadata.

Figure 3.1 shows the structure of the classes and attributes more in detail.

There are different ways to use DCAT. Usually it is used with RDF. At first, a catalog is created and metadata, such as identifier, publisher, theme taxonomy, rights, policies and relevant dates, is added to the catalog. A catalog is of type dcat:Catalog. “dcat:” is a prefix that stands for `http://www.w3.org/ns/dcat#`.

Nowadays, data can be distributed in many ways, such as csv, spreadsheets, etc.. The distribution class helps to express in which way a dataset or a data service is provided. After the creation of a catalog, its content needs to be specified. The content could be datasets, data services and distributions. Datasets are of type dcat:Dataset. To specify the content the identifier for dcat:Datasets needs to be added to the dcat:Catalog. A dcat:Distribution could be used as a subset of a dcat:Dataset.

A subset of a dataset could be, for example, the data of the dataset only in a specific language. Often, a dcat:Distribution can be accessed through a dcat:DataService. Dcat:DataService contains the same metadata as dcat:Dataset. The difference is that the dataservice has an URL endpoint. The class CatalogRecord is used when there is additional metadata. It is not a mandatory class but it can be useful sometimes.

DCAT also provides a way to express data quality by using W3C Data Quality Vocabulary. The option to provide data versioning is still being developed and improved and will probably be fully available in the next upcoming version of DCAT.

3.1.2 Data Privacy Vocabulary (DPV)

The Data Privacy Vocabulary[1] was published by the Data Privacy Vocabulary Community Group. It contains vocabulary to express processing of personal data. The main focus of the ontology is on the following terms:

- **Personal Data Categories**: There exists different type of personal data. Race, religion or politcal opinions are just a few of them. DPV defines the following personal data categories as sub-classes of the class *dpv:PersonalDataCategory*: Interior, External, Social, Financial, Tracking, Historical, Derived Personal Data, and a Special Category of Personal Data.
- **Purposes**: To indicate the purpose of data processing is a requirement of the GDPR and therefore also represented in DPV.
- **Processing Categories**: Processing data means to elaborate and use data. DPV defines the class *dpv:Processing* with various sub-classes for different types of data processing.
- **Technical and Organisational Measures**: These are measures taken to protect data and to comply with legal obligations.
- **Legal Basis such as Consent**: These terms serve to describe if data subjects have given their consent to use their data.
- **Entities such as Recipients, Data Controllers, Data Subjects**: These terms are used to describe stakeholders of the data.
- **Rights**: This term refers to all the rights applicable, provided, or expected by data subjects, data controllers, or other stakeholders.
- **Risks**: The term Risks serves to represent undesirable events or negative impacts of the data.[6]

The authors also defined some use cases such as:

1. represent policies for personal data handling
2. represent information about consent e.g. provenance of consent
3. log/document personal data handling actions e.g. by a data controller
4. check compliance after processing[6]

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

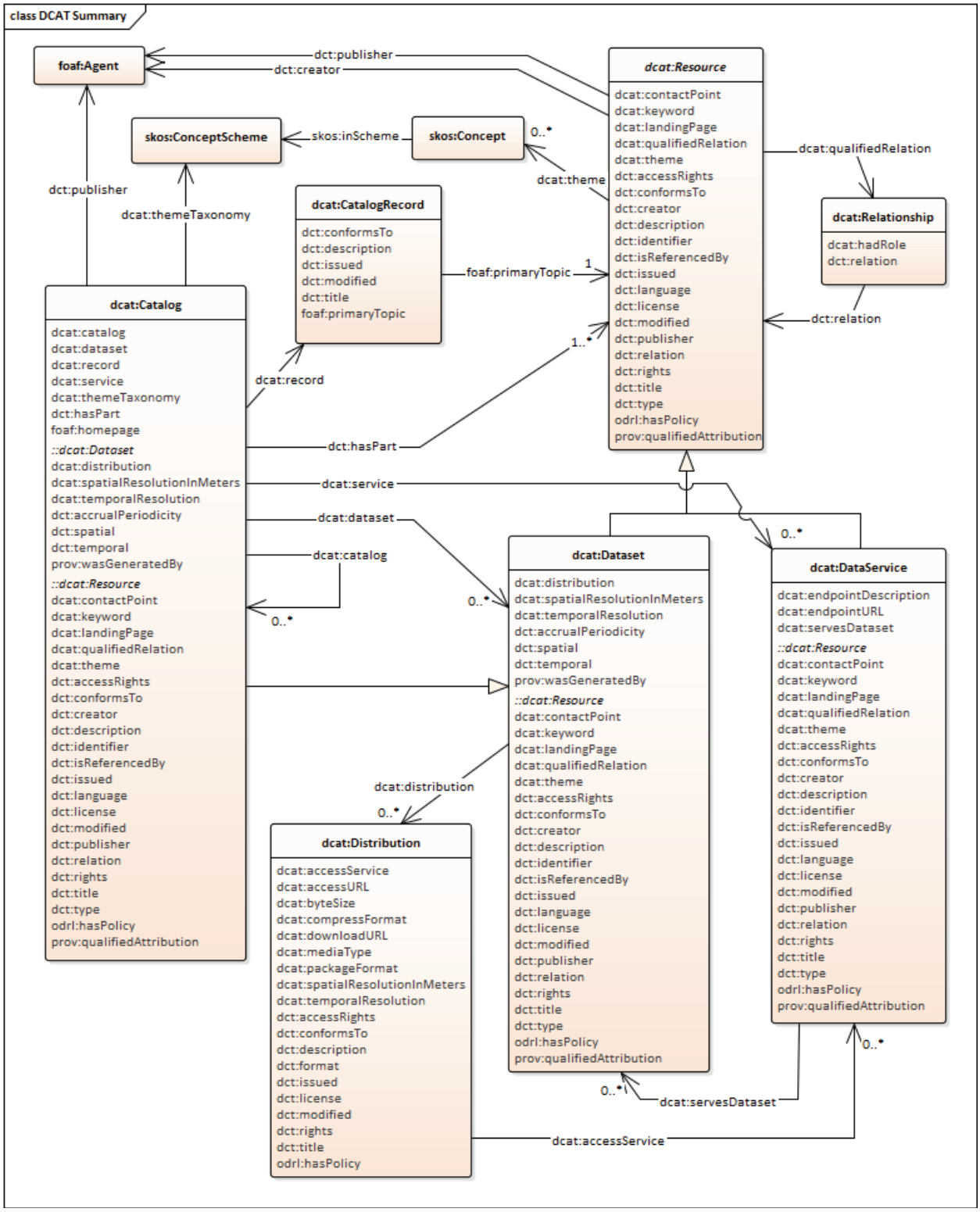

Figure 3.1: DCAT Classes and Attributes [5].

Since the ontology is heavily related to the ontology of this work, there are some classes that can be reused. In specific, classes like DataSubject, DataController, Risk, Purpose, Processing, and LegalBasis will be of particular interest to us. Figure 3.2 shows the main classes of DPV.

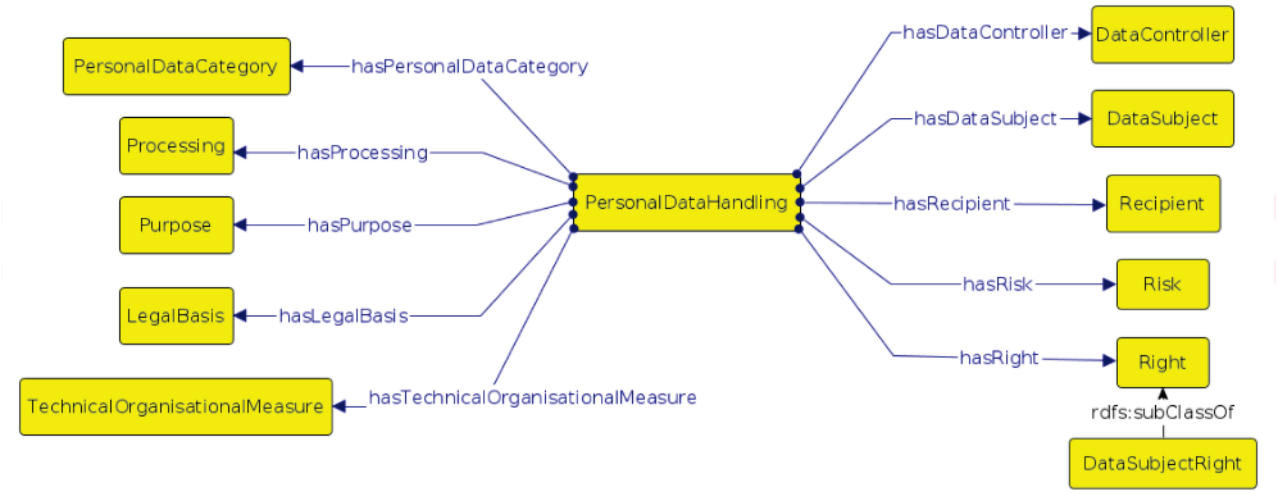

Figure 3.2: DPV main classes [6].

3.1.3 Data Quality Vocabulary (DQV)

The Data Quality Vocabulary is an extension of the DCAT vocabulary. It provides the possibility to provide metadata of datasets in terms of data quality. Since *quality* is an abstract term, the goal of DQV is to enable data publishers to provide enough information about the dataset so that potential users of the dataset can make up their own mind about the quality. The classes that can be reused for the ontology of this project are *QualityMeasurement* and *Metric*. These are the most basic classes of the ontology and allow to provide enough information about data quality for the scope of this work. The main structure of the Data Quality Vocabulary is shown in Figure 3.3.

3.1.4 Comparison of Related Ontologies

The previously presented ontologies that are related to the domain of the project cover different aspects of the domain and are overlapping in some areas with the project that will be described in this thesis. Data Catalog Vocabulary (DCAT), for example, covers general metadata about the dataset as well as some legal aspects with regards to it. On the other hand, Data Privacy Vocabulary (DPV) deals strictly with legal issues. Since Data Quality Vocabulary (DQV) deals with data quality, it can be claimed that it covers indirectly social and ethical impact of the data. On the one hand, low data quality can have a bad impact while, on the other hand, high quality could have a positive social or ethical impact. Table 3.1 summarizes the areas covered by each of the previously presented ontologies.

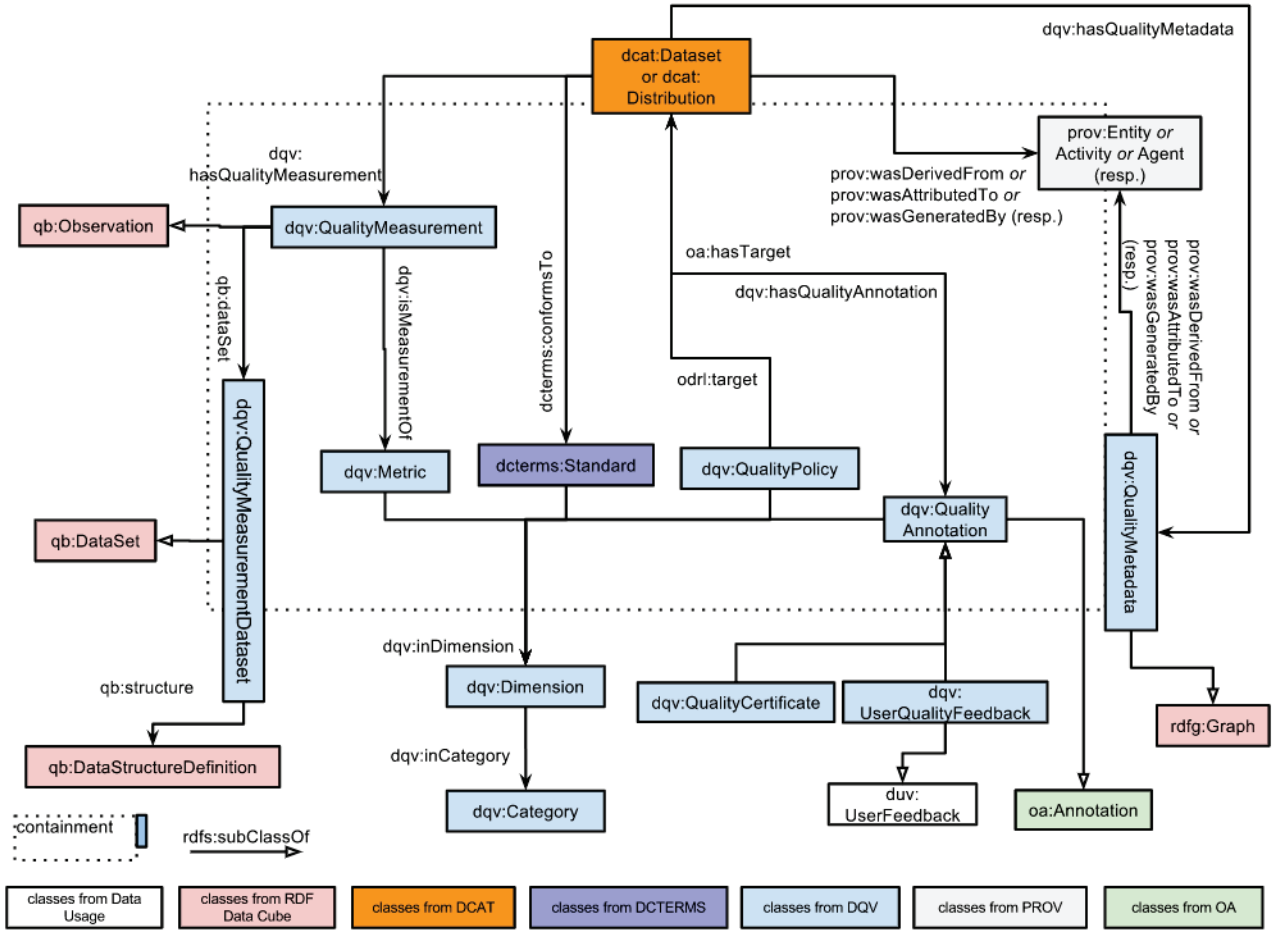

Figure 3.3: DQV main structure [7].

Ontology	General Metadata	Legal Compliance	Social/Ethical Impact
DCAT	**Yes**	**Yes**	No
DPV	No	**Yes**	No
DQV	No	No	**Yes**

Table 3.1: Topics covered by ontologies

3.2 Ethical Assessment. Guidelines and Templates

3.2.1 The Ethics Canvas

The Ethics Canvas was created by ADAPT[2]. ADAPT is an organization based in Ireland and consists of a group of researchers with a common goal which is to reach a "Balanced Digital Society" by 2030. They are researching in different areas, such as AI, natural language processing, data analytics, intelligent machine translation, and human-computer interaction. Their goal is also to impose a standard for data governance, privacy and ethics for digital content. The idea behind the Ethics Canvas is to provide an easy way to assess the ethical impact of a project. According to ADAPT, the Ethics Canvas should be used directly by researchers and developers of the project.

Also, It is a collaborative tool allowing people to work together easily as ADAPT thinks ethical assessment should be conducted by a group of people rather than an individual alone. The canvas can be used either online through the website[3] or offline by printing it out.

The canvas consists of eight fields:

1. Individuals affected
2. Groups affected
3. Behaviour
4. Relations
5. Worldviews
6. Group Conflicts
7. Product or Service Failure
8. Problematic Use of Resources

The first two fields describe all the parties that will be affected by the project. The third one, *Behaviour*, should be used to think about problematic behavioral changes of the affected individuals. On the other hand, in the field *Relations*, researchers and developers of the project can write all the changes in habits and choice of activities of the affected individuals. The field *Worldviews* is defining the impact of the project on society and the world in general. *Group Conflicts* is a field to describe all the groups that could get into a conflict. *Product or Service Failure* should be used to describe all the possible product and service failures and their potential impact. Last but not least, all the potential problematic consumption of resources can be mentioned in the field *Problematic Use of Resources*.

Filling out all these fields allows researchers and developers to get a quick overview of the ethical impact of their project.

[2] https://www.adaptcentre.ie
[3] https://www.ethicscanvas.org

3.2.2 DPIA Template

DPIA is the acronym of Data Protection Impact Assessment. The General Data Protection Regulation (GDPR) that was imposed by the European Union in May 2018 specifies that DPIA has to be conducted for every high-risk processing activity. GDPR does not specify what should exactly be in the DPIA and does not provide a strict template for it. However, in article 35 (7) it specifies that it should at least contain the following information:

- a systematic description of the envisaged processing operations and the purposes of the processing, including, where applicable, the legitimate interest pursued by the controller;
- an assessment of the necessity and proportionality of the processing operations in relation to the purposes;
- an assessment of the risks to the rights and freedoms of data subjects referred to in paragraph 1; and
- the measures envisaged to address the risks, including safeguards, security measures and mechanisms to ensure the protection of personal data and to demonstrate compliance with this Regulation taking into account the rights and legitimate interests of data subjects and other persons concerned.[8]

The U.K. Information Commissioner's Office (ICO) published a template to carry out a complete DPIA[9]. The template contains seven steps. The first two steps are to identify the data processing involved in the project, its nature, scope, context and purpose. Step 3 is to describe the consultation process. In particular, this step requires to specify how and when individuals' points of view are sought, and in case it is considered unnecessary, a justification is needed. In step 4, the necessity of the project is assessed. Furthermore, the process to guarantee data quality and data minimization has to be explained at this point. Step 5 consists of identifying all the risks and describing their likelihood and impact. In step 6, all the measures taken to reduce or eliminate the risks specified in step 5 need to be specified. Last but not least, in step 7, the document is signed off and its outcomes are recorded.

3.2.3 Datasheets for Datasets

Since there is no standardized process for documenting machine learning datasets, Gebru et al., have published a guideline to document datasets. They call it the "Datasheets for Datasets"[10]. The idea is to follow a set of questions that answer all the important aspects of the dataset. This datasheet is interesting for both data creators and data consumers. By using the datasheet, data creators are forced to reflect on the whole process, from the creation of the dataset to its distribution. On the other hand, data consumers will be able to get all the necessary information before using the dataset. Answering the questions of the datasheet could result in a time-consuming activity. However, the process should not be automated since it should be a reflected process. The set of questions are grouped into sections: motivation, composition, collection process, preprocessing/cleaning/labeling, uses, distribution and maintenance. In the motivation section data creators need to specify the reason for the creation of the dataset. The composition section is for data consumers to get information about the dataset and to help them to decide how to use the data. In this

section the data creators describe what kind of data the dataset contains with many details. This section also contains information about the compliance with the EU's GDPR. The *collection process* section is self-explanatory and is the section where data creators explain how they created the dataset. This facilitates an eventual recreation of the dataset in the future. *Preprocessing/cleaning/labeling* is a section where data consumers see what has been done to the data and whether the processed data is suited for their task. In the section *uses*, data creators provide the use cases of the data. Furthermore, they should also specify what should not be done with the data. *Distribution* is a section that is used to describe how the dataset will be distributed and what kind of access restrictions will be imposed on the data. Last but not least, before distributing the data, data creators should provide information to specify how they have planned the maintenance of the dataset.

Dataset creators can omit some of the questions if the question is not applicable to their case. It could be possible that experts from other domains, such as anthropology, are needed during the creation of the datasheet. Creating datasheets by answering all the questions will facilitate the communication between dataset creators and dataset consumers.

3.2.4 Accountability on the ground Part II: Data Protection Impact Assessments Prior Consultation

The part II of the Accountability on the ground toolkit is a guideline published by the European Data Protection Supervisor (EDPS) to analyze and control the risks of a "high risk" project. The document specifies how to carry out a DPIA, when it is necessary to send DPIAs to EDPS or Data Protection Authorities (DPAs) for prior consultation, who is responsible for the completion of these tasks and the transition Rules from old regulations for EU institutions.

Figure 3.4 illustrates the roles of each individual.

	Responsible	Accountable	Consulted	Informed
Top Management		X		
Business owner	X			
DPO			X	
IT department			X	
Processors, where relevant			X	
Data subject representatives			(X)	

Figure 3.4: Roles of Individuals ([11]).

The business owner is responsible for the DPIA, the top management is the accountable party and other individuals/groups, most importantly the DPO, can be consulted.

EDPS does not impose a standard methodology to carry out DPIA. However, EDPS specifies all the steps that are part of the DPIA process: Description of processing,

Assessing necessity and proportionality, Risk analysis, Risk treatment, Sign-off, and Check and review.

In the first part, the process in general is described. It should be mentioned how the data is collected and why it is collected. In the second section, *Assessing necessity and proportionality*, it is necessary to explain the purpose of the processing. It must be ensured that the advantages outweigh the disadvantages that come from it. In the risk analysis part, all the possible harms that could be caused by the process should be listed. It is also important to specify the impact and likelihood of each risk.

EDPS also specifies guiding questions on data protection principles and categorizes them into the following categories: Fairness, Transparency, Purpose limitation, Data minimisation, Accuracy, Storage limitation and Security. For each category, EDPS suggests questions that should be answered. All the proposed questions should be used as a starting point to eliminate problematic aspects of planned processing operations.

In the risk treatment part, data controllers should describe all the measures taken to reduce, eliminate and counter the risks mentioned in the previous section. It should not only be about mitigation measures but also about possible approaches to minimise risks. Therefore, EDPS groups generic controls into four categories of *preventative*, *detective*, *repressive* and *corrective*.

After all of the mentioned tasks, the outcome should be documented. The main deliverable at the end will be a DPIA report. The DPIA should be reviewed regularly. The higher the number of processing operations the shorter should be the review cycles. If the risks could be minimized following the DPIA but there are still "high residual risks," EDPS should be consulted. Carrying out the whole DPIA process proves that it has been thought proactively of all the risks and problematic aspects.

3.2.5 AI FactSheets 360

To increase transparency and improve data governance, IBM Research created the AI FactSheet 360[4]. The idea of the factsheet is to provide information to data consumers about an AI model or service. This information will allow consumers to better understand whether the model or service is suitable for their case. The factsheet of an AI model or service contains relevant information about the creation and deployment. So, it provides information about what dataset had been used, the purpose of it and all the actions taken during the whole process. To gather all this information it could be necessary to involve various stakeholders, such as the business owner, data scientists, model validator and model operator. Each of them can contribute important information to the factsheet. IBM Research claims that it is not possible to use one factsheet for different AI models or services. Each factsheet of a model or service should be tailored specifically to the particular AI project. For this reason, IBM Research published a template which can be used to create the factsheet. The template can be useful to determine which information is relevant to the particular AI project.

[4] https://aifs360.mybluemix.net

3.2.6 Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment

The European commission published the Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment in July 2020. This list consists of questions that should be answered before carrying out a project. Based on the answers it should be possible to determine whether an AI project can be considered as trustworthy. The questions are divided into seven key requirements:

1. human agency and oversight
2. technical robustness and safety
3. privacy and data governance
4. transparency
5. diversity, non-discrimination and fairness
6. environmental and societal well-being
7. accountability[12]

For each requirement there are a set of questions. The goal is to see whether the AI system is interfering human's autonomy or behavior and to check whether the human has the possibility to correct a decision that an AI system has taken wrongly.

The second category, "Technical Robustness and safety", is self-explanatory and covers the topic of safeness and risks that come with the system.

"Privacy and data governance" deals with data protection. In particular, the questions of this requirement aim to show whether data is being protected and used in a responsible way.

The questions about transparency focus on traceability, explainability, and communication about the limitations of the AI system.

The Diversity, non-discrimination and fairness questions have the purpose to find out whether the AI system acts in a discriminatory way and if the system is accessible to everyone.

"Environmental and societal well-being" deals with the impact of the AI system on the environment, work and society.

Last but not least, the questions about accountability are about risk management and identifying measure to mitigate risks.

Once the questions of the various requirements are answered, a judgment can be made whether the AI system is trustworthy.

3.3 The SIENNA project

The University of Twente carried out the SIENNA project together with other partners. The focus of the project is to address ethical issues with regards to areas related to human genomics, human enhancement and human-machine interaction. After identifying the ethical issues and risks, they created three different frameworks, one for

each of the areas. Starting from these frameworks they developed tools for managing ethical issues with new technologies. The main outcome of the project is various guidelines for social and ethical assessments as well as proposals for regulatory aspects in the different technological areas.

Chapter 4

Background

This section serves to introduce the concept of semantic web including RDF, OWL, ontology competency questions and the Linked Open Terms ontology engineering methodology.

4.1 Semantic Web

The term Semantic Web can be considered as a vision of a web of linked data. It aims to make Internet data machine-readable. The standards are set by World Wide Web Consortium (W3C). To represent metadata, technologies like Resource Description Framework (RDF) and Web Ontology Language (OWL) can be used.

RDF could be compared with Entity–relationship models which describe relationships between different entities. However, RDF stores information as so-called *triples*. The triples consist of subject, predicate and object. The predicate defines a relationship between the subject and the object.

Another important aspect to mention about RDF is the usage of URLs to uniquely identify and define entities, relationships and objects. This not only ensures uniqueness but also allows the reuse of entities and objects.

To serialize RDF triples, languages like Turtle, JSON-LD or XML can be used. A collection or group of connected RDF expressions can be visualized in a graph. Figure 4.1 represents graphically a triple.

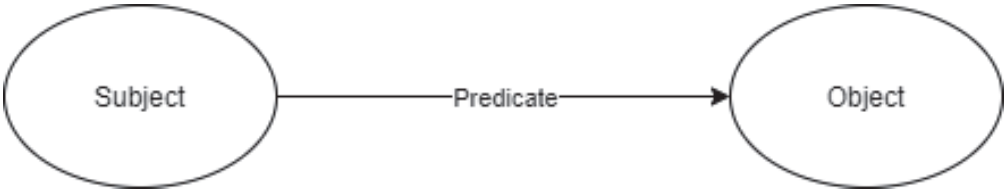

Figure 4.1: RDF triple

On the other hand, OWL is a logic-oriented language to represent knowledge. A knowledge of a domain is called an ontology and can be exploited and edited with specific software.

An ontology consists of classes. Each class can have properties. A property can either be of type object or of type data. Object properties define the relationship between one class with another class while data properties can be thought of attributes of a class. Graphically, classes can be represented as rectangles while object properties can be drawn as arrows that point from one class (the domain) to another class (the range). On the other hand, data objects can be found inside the classes.

Classes can also have sub-classes which inherit the properties from their parent-classes. A subclass has a "kind-of" relationship with its parent-class. In other words, a child is a "kind of" its parent. Fig 4.2 represents a class-hierarchy graphically. Classes can also be instantiated. An instance of a class is the lowest level of granu-

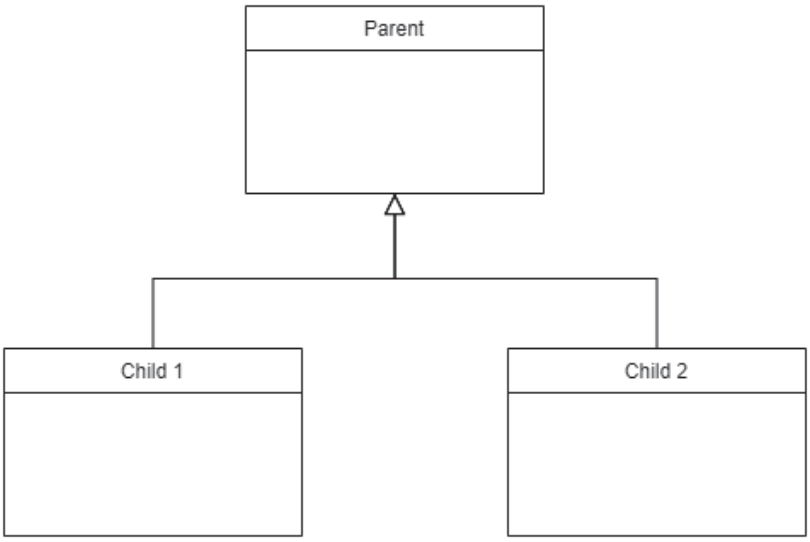

Figure 4.2: Class Hierarchy

larity in the representation. The level of granularity is different for every ontology and depends on the scope and the domain. In general, ontologies in OWL are expressed and stored in RDF.

One possibility, for example, to create and edit an ontology is Protégé[1]. Protégé is an open-source ontology editor and counts with more than 300.000 users[13]. It allows users to either create ontologies from scratch or to edit existing ontologies. Protégé allows to create classes, object properties and data properties. Before starting to create any entities or objects, it is good practice to define the namespace of the ontology. Having a namespace from the beginning means that it does not have to be changed later on since an URI will be assigned to each class and object when it is created. Labels and comments can also be defined. Furthermore, it is possible to reuse ontologies by importing classes and properties.
Once classes and objects have been defined, it is possible to instantiate and use them.

After fully implementing the ontology, the file can be saved in different formats, being RDF and Turtle (ttl) the most prominent solutions, alongside JSON-LD.

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

4.2 Ontology competency questions

Competency questions are used to specify the scope of an ontology and are therefore defined previous to the implementation. In concrete, competency questions are a set of questions that the ontology should be capable of answering. Sometimes the questions are grouped by categories and topics of the domain. The ontology should provide a way to represent enough information so that all the competency questions can be answered. Checking if the ontology can answer all the competency questions is one of the methods for evaluating the completeness of an ontology.

4.3 Ontology Requirements Specification Document (ORSD)

The Ontology Requirements Specification Document (ORSD)[2] is a document that helps to define the requirements for an ontology. In particular, it consists of the following chapters:

- “Purpose” - to define the general goal
- “Scope” - to specify the level of detail and coverage of the ontology
- “Implementation Language” - to specify the formal language of the ontology
- “Intended End-Users” - to define the expected end-users
- “Intended Uses” - to illustrate the intended use
- “Ontology Requirements”
  - “Non-Functional Requirements” - to define the requirements to fulfill
  - “Functional Requirements: Groups of Competency Questions” - to define questions that the ontology should give answers to
- “Pre-Glossary of Terms”
  - “Terms from Competency Questions” - to list the terms and their frequencies of the competency questions
  - “Terms from Answers” - to list the terms and their frequencies of the answers of the competency question
  - “Objects” - to list the objects of the competency questions and their answers

4.4 Ontology engineering method

Ontology engineering methodologies are well-defined procedures for creating ontologies. They define the whole process of creating an ontology. Kotis et al states that ontology engineering methodologies are a necessity in order to obtain a "‘shared, commonly agreed and continuously evolved ’live’ conceptualizations’ of domains of discourse" [14].

[2] https://github.com/oeg-upm/ORSD-template

4.4.1 Linked Open Terms Methodology (LOT)

The Linked Open Terms Methodology is an ontology engineering method to create and public ontologies. The methodology consists of four main parts:

1. Ontology requirements specification
2. Ontology implementation
3. Ontology publication
4. Ontology maintenance

During the ontology requirements specification, the ontology creators define the use cases, purpose and scope of the ontology.
Then, they specify the functional ontology requirements which are all the demands that the ontology should fulfill. Often, they are defined in the form of competency questions.

Once these steps are completed the ontology can be implemented. During the ontology implementation phase the ontology is conceptualized, encoded, and evaluated. It is also possible to reuse other ontologies during this process.

After the ontology is created, it is time to prepare it for publication. Therefore, the documentation is carried out. This is often done by creating an HTML document. Once the documentation is ready the ontology can be published.

The last step, *maintenance*, consists of detecting and fixing bugs and, in case of necessity, meeting new requirements. Figure 4.3 shows the workflow, as well as the various actors for each step in the workflow.

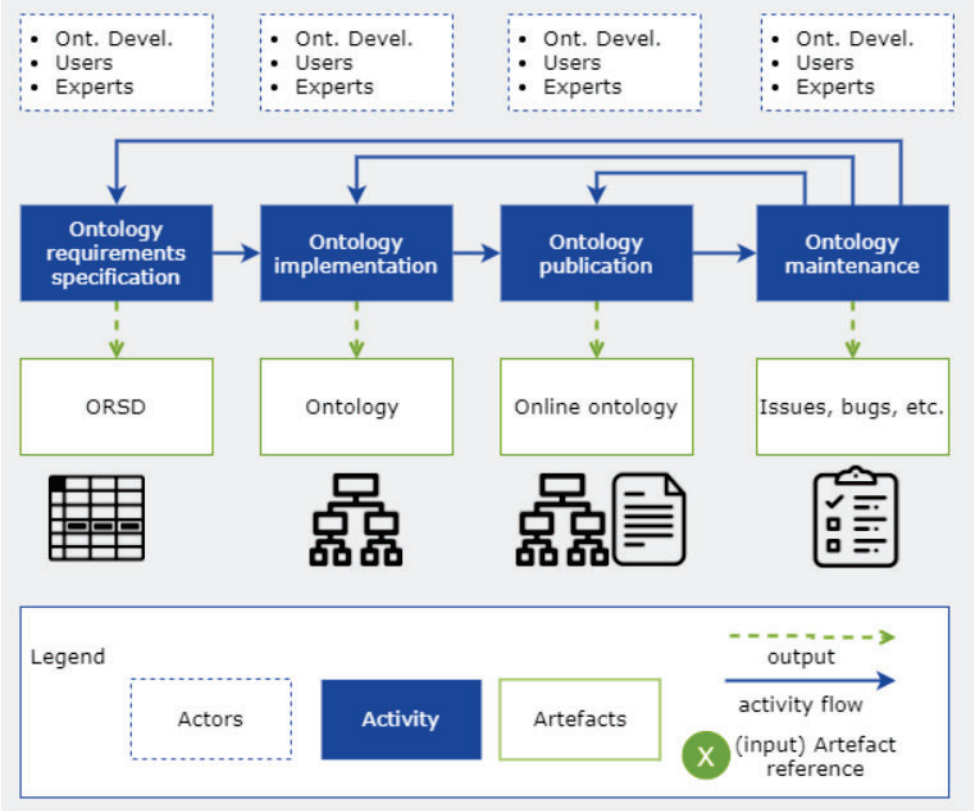

Figure 4.3: Linked Open Terms Methodology Workflow[15].

Chapter 5

Development

This chapter describes the whole process of building the ontology. The previously described ontology engineering methodology, Linked Open Terms Methodology (see 4.4.1), was followed to create the ontology. Therefore, this chapter is divided into four sections, one for each part of the ontology engineering process respectively.

5.1 Ontology requirements specification

Step 1 in the LOT methodology is to define purpose, scope, and use cases for the ontology. It has been decided that the purpose is to create a data model that can be used to describe datasets in terms of their legal, ethical and social impact. The scope is the domain of legal and ethical assessment using DPIA and ethical assessments.

For the use cases it is important to define both, the intended end-users and the intended uses. The following end-users were defined for this ontology:

- Public and private institution that want to publish data on a data portal
- Software Developers and data scientists that want to use a dataset in their projects
- Data Supervisory Authorities

Once the intended users were defined, the here-listed use-cases were elaborated:

- Public authorities who wish to define the legal, social and ethical impact of a dataset when they publish it on a data portal.
- Software Developers and data scientists evaluate together with their teams whether the dataset is appropriate for their project and assess the possible risks associated with the dataset.
- Data Supervisory Authorities (DPAs) want to check the content of a dataset.

After the use-cases had been elaborated, the competency questions of the ontology were defined. The competency questions cover information to identify the collectives or communities that can be affected by processing the data, such as environmental and religious groups, unions, professional bodies, competing companies and government agencies. Furthermore, they include questions about the type of data subjects,

data quality, purpose of data collection and processing, whether and how ethical assessments were carried out. They also identify the risks of processing the data, as well as if measures were taken to avoid or reduce the likeliness of undesirable events. Moreover, there were defined concepts and properties that relate to the competency questions to facilitate the development of the ontology later on. Figure 5.1 shows the concepts table which includes the complete list of competency questions alongside with the concepts that were defined for the ontology:

CQ	Questions	Concepts	Properties
CQ1	What was the purpose of collecting the data?	Data	purposeOfCollectingData
CQ2	What is the size of the dataset?	Data	size
CQ3	How was the data collected? What are the data sources?	Data Source	hasDataSource
CQ4	Who are the data subjects?	Data Subject	hasDataSubject
CQ5	Does the dataset contain personal data?	Data	isPersonalData
CQ6	What type of personal data does it contain?	Data	typeOfData
CQ7	Does it contain sensitive data?	Data	isSensitiveData
CQ8	Is data anonymized or pseudonymized?	Secury Measure	hasSecurityMeasure
CQ9	Do some of the data subjects belong to the group of vulnerable data subjects?	Data Subject	isVulnerable
CQ10	How was data quality ensured?	dqv:QualityMeasurement	dqv:hasQualityMeasurement
CQ11	What are the possible future applications for the dataset? What other taskscould the dataset be used for?	Processing	involvedIn
CQ12	What are the possible risks of using the dataset?	Risk	hasRisk
CQ13	Which collectives or communities, e.g. groups or organisations, can be affectedby the potential use of this dataset?	Group	affectsGroup
CQ14	Are there possible applications that the data should not be used for?	Discouraged Purpose	hasPurpose
CQ15	What is the legal basis for data collection in case that data was collected from individuals?	Legal Basis	hasLegalBasis
CQ16	What security measures were taken to avoid misuse of data?	Security Measure	hasSecurityMeasure
CQ17	What security measures were taken to mitigate risks?	Risk-mitigation Measure	hasMitigationMeasure
CQ18	Is the data biased? If yes, what kind of bias does it contain?	Bias	containsBias/typeOfBias
CQ19	Were authorities asked for consultation? What was their response?	Authority Evaluation	consultsWith producesEvaluation
CQ20	Has a DPIA been conducted? Who carried it out? Which guidelines or templates were used?	Assessment OrganizationalStakeholder Guideline	conductsAssessment followsGuideline
CQ21	Has an ethical assessment been performed? Which stakeholders were involved in carrying out the ethical assessment? Which standards or guidelines were used? What was the outcome?	Assessment OrganizationalStakeholder Guideline Outcome	conductsAssessment followsGuideline producesOutcome

Figure 5.1: Concepts Table for Ontology

Having defined the purpose, scope, and competency questions for the ontology a last step was needed before starting to implement the ontology. The last step consisted in filling in the Ontology Requirements Specification Document. The description and explanation for each part of the document can be found in section 4.3. Unlike the original template suggests it was decided to not define the terms from the answers of the competency questions. This decision was taken due to the special type of competency questions of the domain. Since most of the question are aimed to get specific information about a dataset they often cannot be answered in a general way and therefore it would not have been reasonable to provide default answers to the questions. Table 5.1 shows the final document.

<table>
<tr><th colspan="4">WICUS Ontology Requirements Specification Document</th></tr>
<tr><th colspan="4">1. Purpose</th></tr>
<tr><td colspan="4">The purpose of the Legal, Ethical and Social Impact Ontology is to create a data model that can be used to describe datasets in terms of their legal, ethical and social impact.</td></tr>
<tr><th colspan="4">2. Scope</th></tr>
<tr><td colspan="4">The ontology focuses on the domain of legal and ethical assessment of data using DPIA and ethical assessments.</td></tr>
<tr><th colspan="4">3. Implementation Language</th></tr>
<tr><td colspan="4">The ontology has been implemented in OWL.</td></tr>
<tr><th colspan="4">4. Intended End-Users</th></tr>
<tr><td colspan="4">User 1. Public and private institution that wants to publish data on a data portal
User 2. Software Developers and Data Scientists that want to use a dataset in their projects
User 3. Data Supervisory Authorities</td></tr>
<tr><th colspan="4">5. Intended Uses</th></tr>
<tr><td colspan="4">Use 1. Public authorities define the legal, social and ethical impact of a dataset when they publish it on a data portal.
Use 2. Software Developers or Data Scientists together with their team evaluate whether the dataset is appropriate for their project and assess the possible risks associated with the dataset.
Use 3. Data Supervisory Authorities check what kind of information the dataset contains.</td></tr>
<tr><th colspan="4">6. Ontology Requirements</th></tr>
<tr><th colspan="4">a. Non-Functional Requirements</th></tr>
<tr><td colspan="4">NFR 1. The ontology shall be published online with standard documentation
.</td></tr>
<tr><th colspan="4">b. Functional Requirements: Competency Questions</th></tr>
<tr><td colspan="2"></td><td colspan="2"></td></tr>
<tr><td colspan="2">CQ1. What was the purpose of collecting the data?
CQ2. What is the size of the dataset?
CQ3. How was the data collected? What are the data sources?
CQ4. Who are the data subjects?
CQ5. Does the dataset contain personal data?
CQ6. What type of personal data does it contain?
CQ7. Does it contain sensitive data?
CQ8. Is data anonymized or pseudonymized?
CQ9. Do some of the data subjects belong to the group of vulnerable data subjects?
CQ10. How was data quality ensured?
CQ11. What are the possible future applications for the dataset? What other tasks could the dataset be used for?
CQ12. What are the possible risks of using the dataset?
CQ13. Which collectives or communities, e.g. groups or organisations, can be affected by the potential use of this dataset?
CQ14. Are there possible applications that the data should not be used for?</td><td colspan="2">CQ15. What is the legal basis for data collection in case that data was collected from individuals?
CQ16. What security measures were taken to avoid misuse of data?
CQ17. What security measures were taken to mitigate risks?
CQ18. Is the data biased? If yes, what kind of bias does it contain?
CQ19. Were authorities asked for consultation? What was their response?
CQ20. Has a DPIA been conducted?
Who carried out the DPIA?
How was it realized? Which guidelines or templates were used?
CQ21. Has an ethical assessment been performed?
Which stakeholders were involved in carrying out the ethical assessment?
Which standards or guidelines were used?
What was the outcome? What are the possible negative outcomes?</td></tr>
<tr><th colspan="4">7. Pre-Glossary of Terms</th></tr>
<tr><th colspan="4">a. Terms from Competency Questions + Frequency</th></tr>
<tr><td>Purpose 1
Size 1
Risk 1
Bias 1
Stakeholder 1</td><td>Data source 1
Data subject 2
Legal Basis 1
Authority 1
Outcome 2</td><td>Personal Data 1
Sensitive Data 1
Guideline 2
Ethical assessment 2</td><td>Data quality 1
Application 2
Security Measure 2
Consultation 1</td></tr>
<tr><th colspan="4">c. Objects</th></tr>
<tr><td colspan="4">DPIA</td></tr>
</table>

Table 5.1: Ontology Requirements Specification Document

5.2 Ontology implementation

In this section of the ontology engineering process the focus is on creating the ontology. This is the core part of the project.

Since in the previous part the concepts table has already been created (see 5.1), it was possible to directly start with the creation of an ontology model. To get started, a model was first drawn out using the online diagram drawing tool at `https://www.diagrams.net`. At first, all classes and class hierarchies were defined. Then, data properties and object properties were created and associated with the corresponding classes. After finishing the first draft of the ontology model, elements of other vocabularies were considered for reuse. The following classes were imported from the DQV vocabulary (3.1.3): *QualityMeasurement* and *Metric*. From the DPV vocabulary (3.1.2) the classes *DataSubject*, *DataController*, *Risk*, *Purpose*, *Processing* and *LegalBasis* were reused. Furthermore, the DCAT (3.1.1) vocabulary was also reused as it provided the class *Dataset* to the ontology.

Once the ontology model was completed the ontology had to be implemented. The selected tool for carrying out the task was Protégé. Protégé was not only used to implement the entities and properties, but it was also used to define labels and comments for each of them which is a good practice since it will facilitate the creation of the documentation.

The final step of the implementation is the evaluation of the ontology. *Oops!*[1] was used to carry out the evaluation. To use the tool, the serialized ontology (i.e. RDF, OWL, JSON-LD, etc.) has to be imported by the tool. Oops! reads the file and produces a sort of report. The report contains information about pitfalls and errors.

After running the tool for the first time, the following pitfalls were detected:

- Creating unconnected ontology elements
- Missing annotations
- Missing disjointness
- Inverse relationships not explicitly declared
- Using a miscellaneous class

Oops! also classifies the pitfalls from minor to critical. Since none of the detected pitfalls were critical there has not been made any change to the ontology.

Another metric to analyze and evaluate an ontology is to see whether the ontology complies with the FAIR principles[16]. FAIR stands for *Findable*, *Accessible*, *Interoperable*, and *Reusable*. Since unique URIs were used and the ontology will be published online with metadata the ontology follows the prinicple of *Findable*. The ontology is also *Accessible* as it was published using the standard HTTP protocol. The principle of *Interoperability* is also followed since RDF was used for knowledge representation. To reuse the ontology (*Reusability*) is also possible because the ontology provides metadata about classes and objects which facilitates the process of reusing them.

Finally, the last step to evaluate the ontology, was to check whether the initially defined competency questions could be answered by the ontology. After going through

[1] `http://oops.linkeddata.es/`

every question it was decided that the ontology passed the evaluation and meets all the requirements. Therefore, the next step of the Linked Open Terms Methodology, namely to public the ontology, could be taken.

5.3 Ontology publication

Publishing the ontology consists of creating the documentation for the ontology and making the ontology publicly available. To produce the documentation the tool Wizard for Documenting Ontologies (Widoco[2]) was used. Widoco reads the metadata of the ontology and creates an HTML document out of it. The metadata, in this case, were all the entities with their properties as well as the previously defined labels and comments. Once the HTML document was created the ontology had to be made publicly available. Widoco does not only create the HTML document but it also prepares a folder with all the necessary files to public an ontology. The folder includes the just mentioned HTML file, the ontology in various formats (json, nt, ttl, xml) as well as other relevant files. All the files were deployed and the ontology is availabe at the following URL: `https://protect.oeg.fi.upm.es/def/lseid`.

5.4 Ontology maintenance

Ontology maintenance which deals with fixing errors and meeting new upcoming requirements is a topic that will be discussed in the chapter of future work (7.2).

[2] `https://dgarijo.github.io/Widoco/`

Chapter 6

Results

This chapter is focused on presenting and discussing the results. The most important aspect to discuss is the final created ontology. The chosen name for the ontology is: "Legal, Social, and Ethical Impact of Datasets Vocabulary", being "LSEID" the acronym for it. Furthermore, the following namespace was used:

- ns: <http://www.w3.org/2003/06/sw-vocab-status/ns>
- owl: <http://www.w3.org/2002/07/owl>
- dqv: <https://www.w3.org/ns/dqv>
- xsd: <http://www.w3.org/2001/XMLSchema>
- skos: <http://www.w3.org/2004/02/skos/core>
- rdfs: <http://www.w3.org/2000/01/rdf-schema>
- dpv: <http://www.w3.org/ns/dpv>
- art-37: <https://eur-lex.europa.eu/eli/reg/2016/679/art_37>
- par-7: <https://eur-lex.europa.eu/eli/reg/2016/679/art_4/par_7>
- rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns>
- terms: <http://purl.org/dc/terms>
- usage-policy: <http://www.specialprivacy.eu/langs/usage-policy>
- dcat: <http://www.w3.org/ns/dcat>
- vann: <http://purl.org/vocab/vann>
- par-1: <https://eur-lex.europa.eu/eli/reg/2016/679/art_4/par_1>
- lseid: <https://protect.oeg.fi.upm.es/def/lseid>

To get a general overview of the ontology, Figure 6.1 shows the whole structure of the ontology that has been created. The ontology consists of 27 classes. Given the domain, the class *Data* could be considered as the main class of the ontology. It contains basic information about the data such as whether it is personal and/or sensitive data, and also what the type, size, and purpose of the data is. To indicate to which dataset the data belongs, the object property *containsData* with domain *dcat:Dataset*

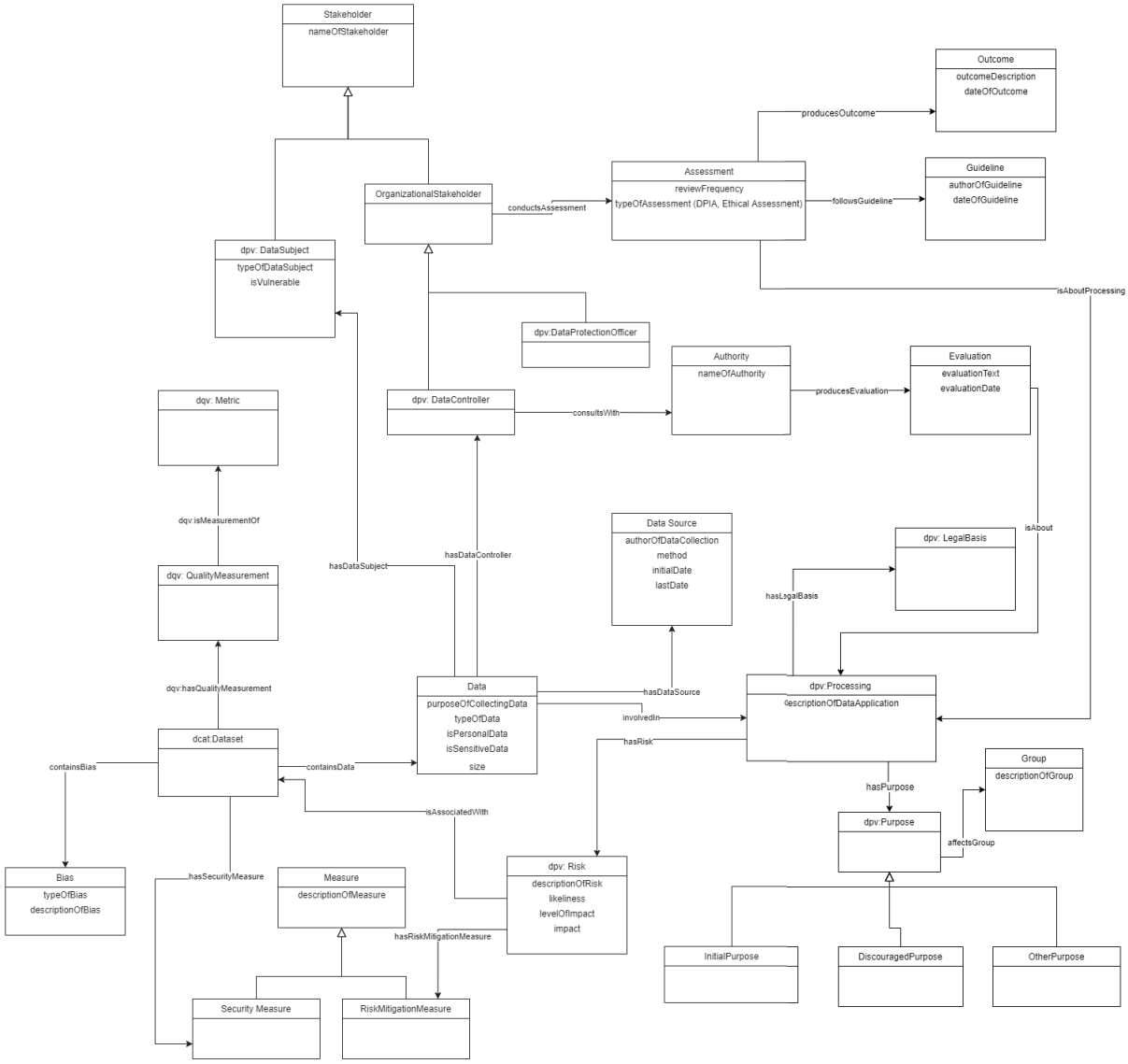

Figure 6.1: Legal, Social, and Ethical Impact of Datasets Vocabulary

and range *Data* has been defined. Since data quality can be a factor that influences the impact of the dataset on society, the class *dqv:DataQualityMeasurement* was reused and connected through an object property with *dcat:Dataset*. In particular, *dqv:DataQualityMeasurement* is used to describe the outcome of measuring the quality of the data. On the other hand, *dqv:Metric* is used to indicate the applied metric to measure the quality of the data. The class *Bias* is also an entity that is connected to *dcat:Dataset*. A bias means that the dataset contains unbalanced data. An example could be when 80% of the collected data is about males while females are only represented in 20% of the data. To describe where the data comes from, the class *DataSource* has been created. This class can be used to indicate where the data comes from, when it was collected ,and who carried out the collection.

The ontology also includes class-hierarchies. As discussed in the previous chapters, in particular chapter 4.1, child-classes are a "kind-of" their super-class. An ex-

ample in the ontology is the class *Stakeholder*. The sub-classes of the Stakeholder class are *dpv:DataSubject* and *OrganizationalStakeholder*. Both satisfy the "kind-of" criteria. The stakeholder entity is used for representing involved parties such as data subjects, data controllers or data protection officers. Since data protection officers are a kind of organizational stakeholder, the corresponding class was defined as a subclass of *OrganizationalStakeholder*. *dpv:DataController* indicates who is controlling the data. In other words, the data controller is the person or a group of people who control and are responsible of the data. *dpv:DataSubject*, on the other hand, can be used to indicate the data subjects which are the affected people whose data was collected.

Another important class is the class *dpv:Processing*. This class represents processing performed on data. The connection, object property, with *Data* indicates which data is being processed. Data processing can be any kind of activity or operation where data is being collected or used. The activities could be, for example, usage of the data for an application, for a data science project, or for an AI system. Processing activities can have impact on society, thus raise ethical questions. For this reason, on one hand, a class to represent the risks (*Risk*) has been created, and on the other hand, a class to illustrate the ethical assessments that were carried out was also created (*Assessment*). *Risk* can be used to indicate a specific risk that comes with processing the data. To give data publishers the opportunity to show how the risks are mitigated, the class *RiskMitigationMeasure* has been created. *RiskMitigationMeasure* is a subclass of *Measure* which is a more general class to represent measures taken to avoid inconvenient events. The other subclass of *Measure* is *SecurityMeasure* which can be used to represent all the security measures that were taken on a dataset. The parent-child relationship of *Measure* with *SecurityMeasure* and *RiskMitigationMeasure* can be seen in Figure 6.2.

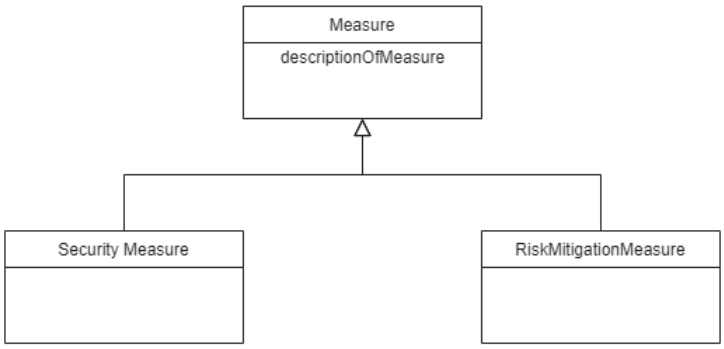

Figure 6.2: Measure Class Hierarchy

While *Risk* deals with inconvenient events, *Assessment* indicates which ethical assessment has been carried out to evaluate the data processing on an ethical perspective. The assessments can be one of the previously described ones in section 3.2, or any other assessments that deal with social and ethical impact. *Assessment* has a connection to *OrganizationalStakeholder* to indicate who carried it out. It is also connected to *Processing* to illustrate to which data processing it is referred. Assessments are usually conducted based on specific guidelines. For this reason, the class is con-

nected to *Guideline*. Finally, the class *Outcome* can be used to publish the outcome of the ethical assessment. Figure 6.3 shows how the classes *OrganizationalStakeholder*, *Assessment*, *Guideline* and *Outcome* are connected with each other.

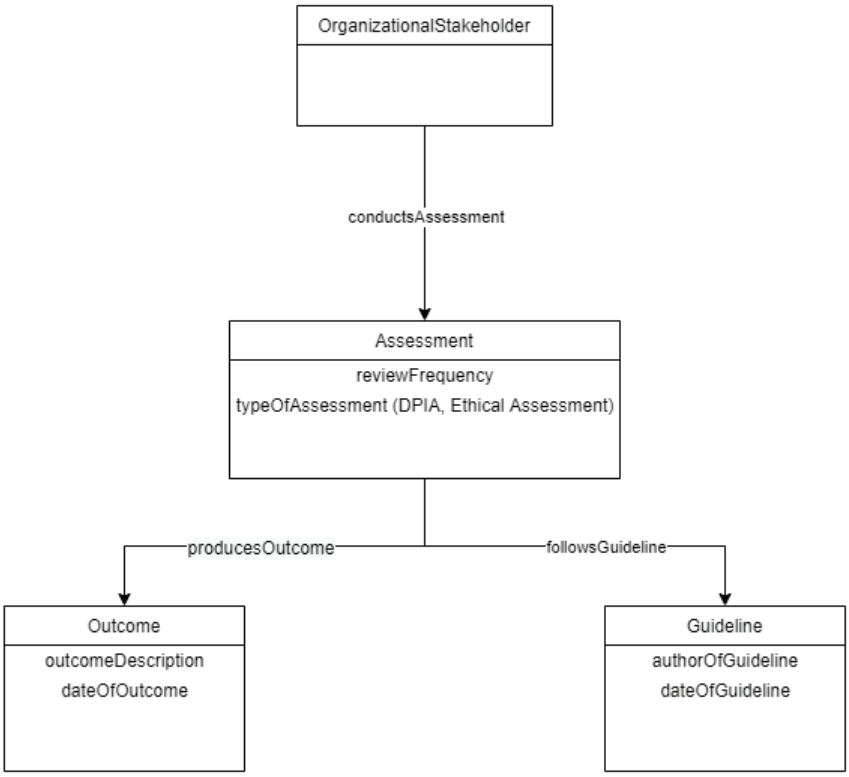

Figure 6.3: Knowledge Representation for Ethical Assessment

To address one of the legal aspects of data processing, the class *dpv:LegalBasis* has been reused. This class can be used, for example, to indicate whether data subjects have given their consent to collect and process their data for specific purposes.

dpv:Purpose is a class to represent the purpose of the data processing. It is the superclass of *InitialPurpose*, *DiscouragedPurpose* and *OtherPurpose* which are all different types of purposes. *InitialPurpose* indicates the purpose that had been defined previously when the data was collected. *OtherPurpose* can be used for secondary purposes and *DiscouragedPurpose* represents discouraged data processing applications which are considered to be dangerous and potentially damaging to society or environment. Figure 6.4 shows the class hierarchy for the class *dpv:Purpose*.

Once the ontology has been created the file has been saved as an OWL file. Figure 6.5 shows a snippet of the produced output.

As described in the previous chapter, once the ontology was implemented the documentation was created and published on `https://protect.oeg.fi.upm.es/def/lseid`. The ontology can be downloaded in the header of the website. It is available as JSON LD, RDF/XML, N Triples and TTL. Furthermore, the header indicates the

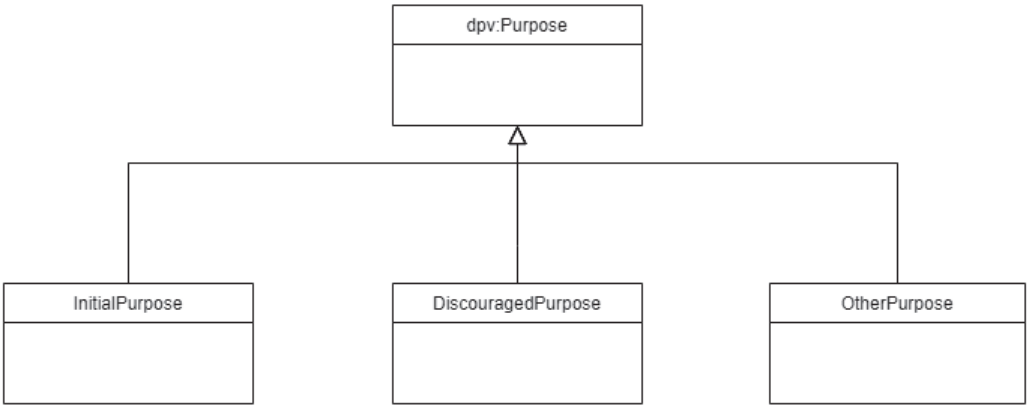

Figure 6.4: Class Hierarchy for the purpose of data processing

chosen license for the ontology which is the creative commons license *CC-BY 4.0*[1]. Figure 6.6 shows the header of the website.

Below the download option a short abstract resumes the ontology. Then, the namespace is introduced. Next, all the entities with their properties are listed and described. In particular, the online documentation displays all the classes and properties, as well as their definition which were specified during the development phase in Protégé. Figure 6.7 illustrates the descriptions for the classes *Assessment* and *Authority*.

[1] `https://creativecommons.org/licenses/by/4.0`

```
###  https://protect.oeg.fi.upm.es/def/lseid#containsData
<https://protect.oeg.fi.upm.es/def/lseid#containsData> rdf:type owl:ObjectProperty ;
                                                       rdfs:domain <http://www.w3.org/ns/dcat#Dataset> ;
                                                       rdfs:range <https://protect.oeg.fi.upm.es/def/lseid#Data> ;
                                                       rdfs:comment "Indicates the content of the dataset" ;
                                                       rdfs:label "containsData" .


###  https://protect.oeg.fi.upm.es/def/lseid#followsGuideline
<https://protect.oeg.fi.upm.es/def/lseid#followsGuideline> rdf:type owl:ObjectProperty ;
                                                           rdfs:domain <https://protect.oeg.fi.upm.es/def/lseid#Assessment> ;
                                                           rdfs:range <https://protect.oeg.fi.upm.es/def/lseid#Guideline> ;
                                                           rdfs:comment "Indicates what Guideline was used for carrying out an Assessment" ;
                                                           rdfs:label "followsGuideline" .


###  https://protect.oeg.fi.upm.es/def/lseid#hasDataController
<https://protect.oeg.fi.upm.es/def/lseid#hasDataController> rdf:type owl:ObjectProperty ;
                                                            rdfs:domain <https://protect.oeg.fi.upm.es/def/lseid#Data> ;
                                                            rdfs:range <http://www.w3.org/ns/dpv#DataController> ;
                                                            rdfs:comment "Indicates the Data Controller of the Data" ;
                                                            rdfs:label "hasDataController" .


###  https://protect.oeg.fi.upm.es/def/lseid#hasDataSource
<https://protect.oeg.fi.upm.es/def/lseid#hasDataSource> rdf:type owl:ObjectProperty ;
                                                        rdfs:domain <https://protect.oeg.fi.upm.es/def/lseid#Data> ;
                                                        rdfs:range <https://protect.oeg.fi.upm.es/def/lseid#DataSource> ;
                                                        rdfs:comment "Indicates where Data was taken from" ;
                                                        rdfs:label "hasDataSource" .


###  https://protect.oeg.fi.upm.es/def/lseid#hasDataSubject
<https://protect.oeg.fi.upm.es/def/lseid#hasDataSubject> rdf:type owl:ObjectProperty ;
                                                         rdfs:domain <https://protect.oeg.fi.upm.es/def/lseid#Data> ;
                                                         rdfs:range <http://www.w3.org/ns/dpv#DataSubject> ;
                                                         rdfs:comment "Indicates who the Data Subjects are" ;
                                                         rdfs:label "hasDataSubject" .
```

Figure 6.5: Turtle File snippet

Figure 6.6: Documentation Header

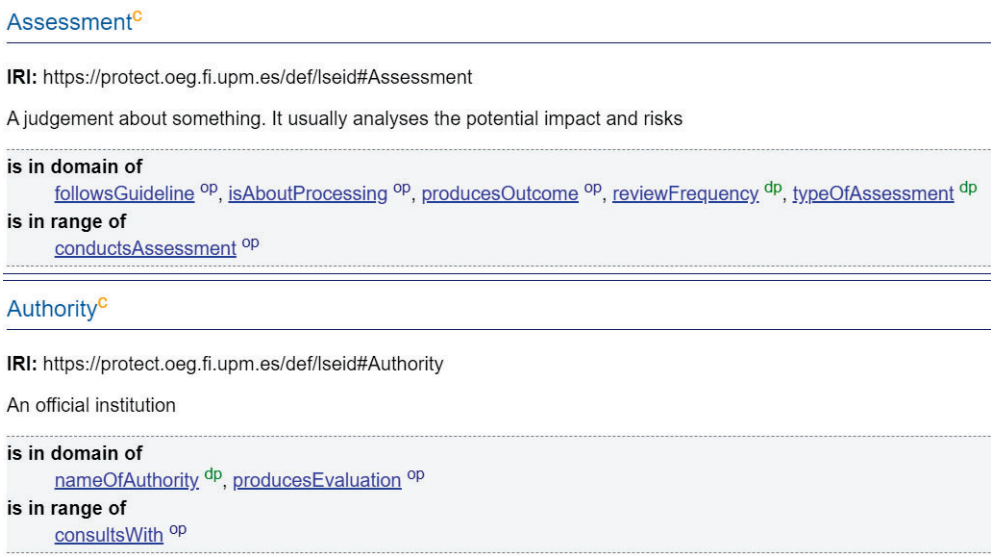

Figure 6.7: Documentation for Assessment and Authority

Chapter 7

Conclusion and Future Work

7.1 Conclusion

The Legal, Social, and Ethical Impact of Datasets Vocabulary created in this project includes all the terms needed to describe datasets in terms of legal, social and ethical impact and therefore reaches the main objective of the project.

The first sub-objective, namely to investigate the required information for assessing legal impacts of a dataset and to create a vocabulary for it has also been met. The focus was set on the Euopean GDPR to cover legal aspects of the research. The class LegalBasis as well as properties like isPersonalData, isSensitiveData guarantee that people interested in a particular dataset can see whether the data publishers considered legal aspects. Also other classes, such as Assessment, Risk, RiskMitigationMeasure are all mentioned concepts in the GDPR, as the Regulation mandates DPIA assessment, having a "risk-based" approach, and having proper security and risk mitigation measures.

The second sub-objective, investigating the necessary information for assessing the social and ethical impact of a dataset and to create a vocabulary for them, has also been addressed successfully. Providing information about bias, ethical assessments that has been conducted on the dataset, as well as risks and purposes, gives potential users of the dataset a complete overview of the social and ethical impact of the data.

The third sub-objective consists of extending the W3C Data Catalog Vocabulary. This sub-objective has also been achieved since the class *dcat:Dataset* was imported and reused in the ontology.

Since all the objective could be addressed successfully, the goal of providing the possibility to data publishers to provide a complete overview of a dataset can be reached. Based on the metadata, data scientists, software developers, project managers or other interested individuals can decide if a particular dataset is a valid option for their data science, software development, or AI project.

Furthermore, another point worth mentioning is that this ontology could become a common set of vocabularies that is understandable by all the involved parties. This is not only beneficial for understanding better the content and risks of a given dataset, but it also improves and facilitates the whole data transmission process between involved parties.

7.2 Future Work

Ontology developments are never-ending projects since ontology maintenance is a continuous process. Governments and legislation constantly change over time. Therefore, it is most likely that in the next years there will be new laws and new aspects to consider when talking about data and data processing. These new requirements may not be covered by the current ontology. Similarly, also ethical views change over time. What in today's society is seen as totally normal, could be seen as unacceptable or unethical in future. Moreover, it is possible that there will be new innovative ways to assess social and ethical impact of data. These new, both legal and ethical, requirements that can emerge in the future need to be addressed.

Another possible future task could be associated with the level of granularity of the ontology. The ontology has been developed on a general level. There is plenty of possibility to go more in detail in legal and also in ethical aspects by, for example, representing different information items in an ethical assessment.

Bibliography

[1] Bernard Marr. How much data do we create every day the mind blowing stats everyone should read, May 2018. https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read, accessed on May 1, 2021.

[2] Nicholas Confessore. Cambridge analytica and facebook: The scandal and the fallout so far, April 2018. https://www.nytimes.com/2018/04/04/us/politics/cambridge-analytica-scandal-fallout.html, accessed on July 3, 2021.

[3] The official portal for European data. Open data and data bias, June 2020. https://data.europa.eu/en/news/open-data-and-data-bias, accessed on July 13, 2021.

[4] ec.europa.eu. Horizon 2020 online manual, 2020. https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/ethics_en.htm, accessed on July 13, 2021.

[5] W3C. Data catalog vocabulary (dcat) - version 2, February 2020. https://www.w3.org/TR/vocab-dcat-2, accessed on May 1, 2021.

[6] Data Privacy Vocabularies and Controls Community Group. Data privacy vocabulary, January 2021. https://dpvcg.github.io/dpv, accessed on June 15, 2021.

[7] W3C. Data quality vocabulary, December 2016. https://www.w3.org/TR/vocab-dqv, accessed on June 20, 2021.

[8] intersoft consulting. Art. 35 gdpr. https://gdpr-info.eu/art-35-gdpr, accessed on May 10, 2021.

[9] The U.K. Information Commissioner's Office (ICO). Sample dpia template, 2018. https://iapp.org/media/pdf/resource_center/dpia-template-v04-post-comms-review-20180308.pdf, accessed on May 18, 2021.

[10] Briana Vecchione Jennifer Wortman Vaughan Hanna Wallach Hal Daumé III Timnit Gebru, Jamie Morgenstern and Kate Crawford. Datasheets for datasets, January 2020. https://arxiv.org/pdf/1803.09010v6.pdf, accessed on May 23, 2021.

[11] European Data Protection Supervisor. Accountability on the ground part ii: Data protection impact assessments prior consultation, February 2018. https://ed

ps.europa.eu/sites/edp/files/publication/18-02-06_accountability_on_the_ground_part_2_en.pdf, accessed on May 3, 2021.

[12] European Commission website. Assessment list for trustworthy artificial intelligence (altai) for self-assessment, March 2021. https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment, accessed on May 5, 2021.

[13] Stanford University. Protégé, 2020. https://protege.stanford.edu, accessed on May 10, 2021.

[14] Dimitris Spiliotopoulos Konstantinos I. Kotis, George A. Vouros. Ontology engineering methodologies for the evolution of living and reused ontologies: status, trends, findings and recommendations, January 2020. https://www.cambridge.org/core/journals/knowledge-engineering-review/article/ontology-engineering-methodologies-for-the-evolution-of-living-and-reused-ontologies-status-trends-findings-and-recommendations/7A2D8D844EE0369C24967E156910AB50, accessed on June 25, 2021.

[15] Raúl García Castro María Poveda Villalón, Alba Fernández Izquierdo. Linked open terms (lot) methodology (version 1.0), January 2019. https://lot.linkeddata.es, accessed on May 5, 2021.

[16] María Poveda-Villalón. Fair principles and semantics on the web: where is the meeting point?, April 2021. https://joinup.ec.europa.eu/collection/oeg-upm/news/fair-ontologies, accessed on July 15, 2021.