

Universidade Federal do Espírito Santo

Patricia Marçal Carnelli Campos

**Designing a Network of Reference Ontologies for the Integration of
Water Quality Data**

Vitória - ES, Brazil
October, 2019

Patricia Marçal Carnelli Campos

Designing a Network of Reference Ontologies for the Integration of Water Quality Data

Dissertação apresentada ao Programa de Pós-Graduação em Informática do Centro Tecnológico da Universidade Federal do Espírito Santo como requisito parcial para a obtenção do título de Mestre em Informática.

Orientador: Prof. Dr. João Paulo Andrade Almeida.

Vitória - ES, Brazil
October, 2019

Ficha catalográfica disponibilizada pelo Sistema Integrado de Bibliotecas - SIBI/UFES e elaborada pelo autor

C198d Campos, Patricia Marçal Carnelli, 1983-
Designing a Network of Reference Ontologies for the
Integration of Water Quality Data / Patricia Marçal Carnelli
Campos. - 2019.
140 f. : il.

Orientador: João Paulo Andrade Almeida.
Dissertação (Mestrado em Informática) - Universidade
Federal do Espírito Santo, Centro Tecnológico.

I. Almeida, João Paulo Andrade. II. Universidade Federal
do Espírito Santo. Centro Tecnológico. III. Título.

CDU: 004



DESIGNING A NETWORK OF REFERENCE ONTOLOGIES FOR THE INTEGRATION OF WATER QUALITY DATA

Patricia Marçal Carnelli Campos

Dissertação submetida ao Programa de Pós-Graduação em Informática da Universidade Federal do Espírito Santo como requisito parcial para a obtenção do grau de Mestre em Informática.

Aprovada em 21 de outubro de 2019:


Prof. Dr. João Paulo Andrade Almeida
Orientador(a)


Prof.^a Dr.^a Monalessa Perini Barcellos
Membro Interno


Prof.^a Dr.^a Maria Luiza Machado Campos
Membro Externo, participação remota

*João Paulo Meneguelli Campos, for being my love
and my greatest supporter.*

*To Vitória Carnelli Campos and Sofia Carnelli
Campos, for making my life lighter and happier.*

*To Paulo Ruy Valim Carnelli and Maria Martha
Marçal Carnelli, for the love, dedication and
values transmitted.*

*To Rebecca Marçal Carnelli and Bianca Marçal
Carnelli, for the love and complicity.*

Acknowledgements

First, I thank God and my family. For the family, this is not an easy time, as we often need to be absent. So, I want to thank my family, who gave me support, affection, patience and understanding.

I also thank the dedication of the professors and other collaborators of the Postgraduate Program in Informatics of the Technological Center of the Federal University of Espírito Santo, mainly those from the NEMO laboratory. They work hard to improve this important instrument for the formation of new teachers, researchers and professionals in the computer field.

In particular, I thank my advisor Prof. Dr. João Paulo Andrade Almeida for the opportunity to work with him, for the exchange of knowledge and for his valuable guidance in the development of this work.

I also thank my colleague Cássio Chaves Reginato for his participation in my academic experience. He shared with me most of the activities that have resulted in this work.

Finally, I thank my friend Evellin Cristine Souza Cardoso for introducing me to my advisor.

This work was partly supported by CNPq (407235/2017-5) and CAPES Finance Code 001 (23038.028816/2016-41).

Abstract

Data semantic heterogeneity poses a significant challenge to integrated environmental data reuse. This challenge can be addressed with the use of ontologies that can provide a common semantic background for data interpretation, supporting meaning negotiation. However, there are some barriers to build ontologies for data integration in complex domains such as the environmental one. A relevant problem is the development of new ontologies disregarding previous knowledge resources such as reference models and vocabularies. To deal with this problem, in this work, we propose a systematic approach for the identification and selection of reusable knowledge resources for building ontologies with the purpose of scientific research data integration. The approach (dubbed CLeAR) follows some principles of the Systematic Literature Review, supporting the search for knowledge resources in the scientific literature. We apply the approach to the environmental domain, focusing on water quality. A total of 543 publications were surveyed. The results obtained provide a set of 75 knowledge resources for the environmental domain, evaluated according domain coverage and some quality attributes. In the case of water quality data, there is an ample spectrum of subject domains covered (including geographical features, spatial coordinates, environmental quality parameters, measurement activities, sampling activities, involved organizations, etc.). None of the knowledge resources on their own covers all aspects required to address the integration of water quality data. In addition, they are not always explicitly related, which makes them unsuitable for data integration in their current form. Because of this, in this work, we propose the design of a network of reference ontologies for the integration of water quality data, based on some of the identified knowledge resources. The proposed ontology network is grounded in the Unified Foundational Ontology (UFO), which provides basic notions of object, relation, property, event, and others necessary to model the environmental domain, besides allowing the analysis and adaptation of the concepts represented by different knowledge resources, in order to enable their integration into the ontology network.

Keywords: Data integration; water quality data; reuse; systematic search; ontology network.

Resumo

A heterogeneidade semântica representa um grande desafio para a reutilização integrada de dados ambientais. Esse desafio pode ser enfrentado com o uso de ontologias que fornecem uma base semântica comum para a interpretação dos dados, apoiando a negociação de significados. No entanto, existem algumas barreiras para a construção de ontologias com o propósito de integração de dados em domínios complexos como o domínio ambiental. Uma delas é o desenvolvimento de novas ontologias sem considerar o reuso de recursos de conhecimento existentes, como modelos de referência e vocabulários. Para lidar com esse problema, nesse trabalho, propomos uma abordagem sistemática para a identificação e a seleção de recursos de conhecimento reutilizáveis na construção de ontologias com o objetivo de integrar dados de pesquisas científicas. A abordagem (denominada CLeAR) segue alguns princípios da Revisão Sistemática da Literatura, apoiando a busca de recursos de conhecimento na literatura científica. Aplicamos a abordagem ao domínio ambiental, com foco em qualidade de água. Foram pesquisadas 543 publicações. Os resultados obtidos fornecem um conjunto de 75 recursos de conhecimento para o domínio ambiental, avaliados de acordo com a cobertura do domínio e alguns atributos de qualidade. No caso de dados de qualidade de água, existe um amplo espectro de domínios envolvidos (incluindo características geográficas, coordenadas espaciais, parâmetros de qualidade ambiental, atividades de medição, atividades de amostragem, organizações envolvidas, etc.). Nenhum dos recursos de conhecimento identificados abrange por si só todos os aspectos necessários para abordar a integração de dados de qualidade de água. Além disso, eles nem sempre estão explicitamente relacionados, o que os torna inadequados para a integração de dados em sua forma atual. Por isso, nesse trabalho, propomos o projeto de uma rede de ontologias de referência para a integração de dados de qualidade de água, com base em alguns desses recursos de conhecimento. A rede de ontologias proposta está fundamentada na Ontologia Fundamental Unificada (UFO), que fornece noções básicas de objeto, relação, propriedade, evento e outras necessárias para modelar o domínio ambiental, além de permitir a análise e a adaptação dos conceitos representados por diferentes recursos de conhecimento, a fim de possibilitar sua integração na rede de ontologias.

Palavras-chave: Integração de dados; dados de qualidade de água; reuso; busca sistemática; rede de ontologias.

List of Figures

Figure 1 - Impact of the mud wave on the Doce River. (A) the river in the Camargos Municipality few days after the disaster, (B) dead fishes nearby the Doce River Park, (C) dead fishes at Governador Valadares, and (D) the Doce River mouth 25 days after the dam burst [13].	16
Figure 2 - Ontology Network Architecture proposed by [28].	19
Figure 3 - Activities performed in the development of this work.	20
Figure 4 - Main Activities in Ontology Engineering extracted from [33].	26
Figure 5 - CLeAR activities.	33
Figure 6 - Language used by the structured resources.	57
Figure 7 - Popularity of structured resources according to the number of identified publications that mention them.	57
Figure 8 - Level of reuse of structured resources according to the number of structured resources that adopt them.	58
Figure 9 - A fragment of UFO-A [24].	71
Figure 10 - A fragment of UFO-A related to Qualities [24][25][26][50].	72
Figure 11 - A fragment of UFO-B [24][25].	74
Figure 12 - A fragment of UFO-C related to agents, objects and normative descriptions [25].	74
Figure 13 - The basic Observation type extracted from [51].	75
Figure 14 - The SamplingFeature core extracted from [51].	77
Figure 15 - The Specimen model extracted from [51].	78
Figure 16 - Conceptual Model of QUDT extracted from [52].	81
Figure 17 - UML Class Diagram for the “Hydro - Physical Waters” conceptual schema extracted from [54].	85
Figure 18 - UML Class Diagram for the “AdministrativeUnit” spatial object extracted from [55].	86
Figure 19 - Fragment of the UML Class Diagram for the “Environmental Monitoring Facilities” conceptual schema related to environmental monitoring facilities extracted from [56].	88
Figure 20 - Fragment of the UML Class Diagram for the “Environmental Monitoring Facilities” conceptual schema related to observations and measurements extracted from [56].	89
Figure 21 - UML Class Diagram for the Coordinate Reference System package extracted from [58].	93
Figure 22 - UML Class Diagram for the Coordinate System package extracted from [58].	94
Figure 23 - Tree view of part of the EnvO related to material terms [61].	96
Figure 24 - Tree view of part of the ChEBI Molecular Structure Ontology extracted from [63].	100
Figure 25 - Architecture of the Network of Reference Ontologies for the Integration of Water Quality Data.	105
Figure 26 - The Core Level Ontologies.	106
Figure 27 - The Material Entity Ontology.	106

Figure 28 - The Spatial Location Ontology.....	107
Figure 29 - The Scientific Research Activity Ontology.....	110
Figure 30 - The Research Activity Ontology.....	111
Figure 31 - The Sampling Ontology.....	112
Figure 32 - The Preparation Ontology.....	112
Figure 33 - The Measurement Ontology.....	113
Figure 34 - The Environmental Monitoring Ontology.....	116
Figure 35 - The Water Quality Ontology.....	117

List of Tables

Table 1 - Inputs, Outputs and Actors of Integration Questions Definition	34
Table 2 - Inputs, Outputs and Actors of Data Sources Selection	35
Table 3 - Inputs, Outputs and Actors of Domain Aspects Identification	36
Table 4 - Inputs, Outputs and Actors of Systematic Search Configuration	39
Table 5 - Inputs, Outputs and Actors of Publications Selection.....	40
Table 6 - Inputs, Outputs and Actors of Structured Resources Identification.....	41
Table 7 - Inputs, Outputs and Actors of Snowballing.....	41
Table 8 - Inputs, Outputs and Actors of Systematic Search Reporting.....	42
Table 9 - Structured Resources Domain Coverage Matrix	43
Table 10 - Structured Resources Quality Attributes Matrix.....	45
Table 11 - Inputs, Outputs and Actors of Structured Resources Analysis	45
Table 12 - Inputs, Outputs and Actors of Structured Resources Classification	46
Table 13 - Inputs, Outputs and Actors of Structured Resources Evaluation.....	46
Table 14 - Integration Questions.....	49
Table 15 - Fragment of a Table from the Renova Foundation Weekly Water Quality Bulletin (04-Feb-2019)	51
Table 16 - Concepts of Water Quality used by Brazilian Organizations	52
Table 17 - Keywords related to Structured Resources	53
Table 18 - Keywords related to Research Domain.....	53
Table 19 - Control Papers.....	54
Table 20 - Publications Inclusion and Exclusion Criteria.....	54
Table 21 - Structured Resources Inclusion and Exclusion Criteria.....	54
Table 22 - Domain Coverage for the Structured Resources on the Water Quality Domain	59
Table 23 - Quality Attributes for the Structured Resources on the Water Quality Domain.....	61
Table 24 - Fragment of the Structured Resources Classification	63
Table 25 - Fragment of the Structured Resources Evaluation.....	65
Table 26 - Relations used to classify knowledge resources elements according to UFO concepts (extracted from [49])	69
Table 27 - Relations between O&M Conceptual Model elements and UFO concepts	80
Table 28 - Relations between the QUDT Ontologies elements and UFO concepts.....	84
Table 29 - Relations between the “Hydro - Physical Waters” conceptual schema elements and UFO concepts.....	90
Table 30 - Relations between the “AdministrativeUnit” spatial object elements and UFO concepts...	91

Table 31 - Relations between the “Environmental Monitoring Facilities” conceptual schema elements and UFO concepts	92
Table 32 - Relations between the Coordinate Reference System UML schema elements and UFO concepts.....	95
Table 33 - Relations between the EnvO Material Terms elements and UFO concepts	98
Table 34 - Relations between the ChEBI Molecular Structure Ontology elements and UFO concepts	101
Table 35 - Correspondences between Material Entity Ontology concepts and EnvO Material Terms elements.....	107
Table 36 - Correspondences between Spatial Location Ontology concepts and knowledge resources reused elements	109
Table 37 - Correspondences between Research Activity Ontology concepts and O&M Conceptual Model elements	111
Table 38 - Correspondences between Sampling Ontology concepts and O&M Conceptual Model elements.....	112
Table 39 - Correspondences between Preparation Ontology concepts and O&M Conceptual Model elements.....	113
Table 40 - Correspondences between Measurement Ontology concepts and knowledge resources reused elements	115
Table 41 - Correspondences between Environmental Monitoring Ontology concepts and Environmental Monitoring Facilities UML Model of INSPIRE elements	116
Table 42 - Correspondences between Water Quality Ontology concepts and knowledge resources reused elements	118
Table 43 - Checking the ontology network elements that answer the integration questions	119
Table 44 - Weights assigned to parameters for WQI calculation extracted from [14].....	120
Table 45 - Checking the ontology network concepts that represent the elements of the data sources to be integrated.....	124

Table of Contents

1	Introduction	14
1.1	Motivation	14
1.2	Context: The Doce River Project	16
1.3	Objectives.....	17
1.4	Approach.....	17
1.5	Structure	21
2	Background.....	22
2.1	Ontologies	22
2.2	Ontology Network.....	23
2.3	Ontology Engineering Methodologies	25
2.3.1	<i>The NeOn Methodology</i>	26
2.3.2	<i>Reuse-Related Gaps</i>	28
2.4	Systematic Literature Review	29
3	The CLeAR Approach.....	32
3.1	Overview of CLeAR Activities.....	32
3.2	Cycle I: Data Integration Requirements Definition	33
3.2.1	<i>Integration Questions Definition</i>	34
3.2.2	<i>Data Sources Selection</i>	34
3.2.3	<i>Domain Aspects Identification</i>	35
3.3	Cycle II: Structured Resources Systematic Search	36
3.3.1	<i>Systematic Search Configuration</i>	37
3.3.2	<i>Publications Selection</i>	40
3.3.3	<i>Structured Resources Identification</i>	40
3.3.4	<i>Snowballing</i>	41
3.3.5	<i>Systematic Search Reporting</i>	41
3.4	Cycle III: Structured Resources Selection.....	42
3.4.1	<i>Structured Resources Analysis</i>	42
3.4.2	<i>Structured Resources Classification</i>	45
3.4.3	<i>Structured Resources Evaluation</i>	46
3.5	Concluding Remarks	46
4	Applying CLeAR to the Water Quality Domain	48
4.1	Definition of the Water Quality Data Integration Requirements	48
4.1.1	<i>Integration Questions for the Water Quality Domain</i>	48

4.1.2	<i>Data Sources to be integrated</i>	49
4.1.3	<i>Water Quality Domain Aspects</i>	50
4.2	Systematic Search for Structured Resources on the Water Quality Domain	53
4.2.1	<i>Configuring the Systematic Search</i>	53
4.2.2	<i>Selecting Publications</i>	55
4.2.3	<i>Identifying Structured Resources</i>	55
4.2.4	<i>Applying Snowballing</i>	55
4.2.5	<i>Reporting the Results of the Systematic Search</i>	56
4.3	Selection of the Structured Resources on the Water Quality Domain	59
4.3.1	<i>Analyzing the Structured Resources</i>	59
4.3.2	<i>Classifying the Structured Resources</i>	62
4.3.3	<i>Evaluating the Structured Resources</i>	63
4.4	Related Work	66
4.5	Concluding Remarks	67
5	Ontological Analysis of the Knowledge Resources Selected for Reuse	69
5.1	The Unified Foundational Ontology	69
5.1.1	<i>UFO-A: An Ontology of Endurants</i>	70
5.1.2	<i>UFO B: An Ontology of Perdurants</i>	73
5.1.3	<i>UFO C: An Ontology of Social Entities</i>	74
5.2	The O&M Conceptual Model	75
5.2.1	<i>Overview of the O&M Conceptual Model</i>	75
5.2.2	<i>Ontological Analysis of the O&M Conceptual Model</i>	78
5.3	The QUDT Ontologies	81
5.3.1	<i>Overview of the QUDT Ontologies</i>	81
5.3.2	<i>Ontological Analysis of the QUDT Ontologies</i>	83
5.4	The INSPIRE Conceptual Model.....	84
5.4.1	<i>Overview of the INSPIRE Conceptual Model</i>	84
5.4.2	<i>Ontological Analysis of the INSPIRE Conceptual Model</i>	90
5.5	The ISO/TC 211	92
5.5.1	<i>Overview of the Coordinate Reference System UML Schema</i>	93
5.5.2	<i>Ontological Analysis of the Coordinate Reference System UML Schema</i>	95
5.6	The Environment Ontology (EnvO).....	95
5.6.1	<i>Overview of the EnvO Material Terms</i>	96
5.6.2	<i>Ontological Analysis of the EnvO Material Terms</i>	97
5.7	The ChEBI Ontology	98
5.7.1	<i>Overview of the ChEBI Molecular Structure Ontology</i>	99

5.7.2	<i>Ontological Analysis of the ChEBI Molecular Structure Ontology</i>	100
5.8	Concluding Remarks	101
6	The Network of Reference Ontologies for the Integration of Water Quality Data	102
6.1	The Ontology Network Development Process	103
6.2	The Ontology Network Architecture.....	104
6.3	The Core Level Ontologies	105
6.3.1	<i>The Material Entity Ontology</i>	106
6.3.2	<i>The Spatial Location Ontology</i>	107
6.3.3	<i>The Scientific Research Activity Ontology</i>	109
6.4	The Domain Level Ontologies	115
6.4.1	<i>The Environmental Monitoring Ontology</i>	115
6.4.2	<i>The Water Quality Ontology</i>	116
6.5	Evaluation of the Ontology Network	119
6.5.1	<i>Verification of the Ontology Network</i>	119
6.5.2	<i>Validation of the Ontology Network</i>	123
6.6	Related Work	125
6.6.1	<i>Models for the Integration of Water Quality Data</i>	125
6.6.2	<i>Models related to Scientific Research Activities</i>	126
6.7	Concluding Remarks	127
7	Final Considerations	130
7.1	Summary of the Work	130
7.2	Applicability of the Work in other Scenarios.....	132
7.3	Limitations and Difficulties	133
7.4	Future Work	133
8	References	135

1 Introduction

1.1 Motivation

Research, management and environmental decision-making involve the systematic collection, interpretation and evaluation of environmental data. Given the high costs involved in producing such data [1], it is no surprising that significant gains can be achieved from data sharing, reuse and integration [2]. However, environmental data are often provided by a variety of sources (such as academic institutions, government agencies, private companies and independent research groups), in different contexts (e.g., scientific research, government actions), and for many purposes (analysis of water quality, air quality, etc.). As a consequence, environmental data are available, when they are, in heterogeneous forms.

Data heterogeneity can occur in terms of syntax, schema or semantics [3]. Syntactic heterogeneity is mainly caused due to the use of different serialization formats and technologies. Schematic heterogeneity occurs when data sources use different schemas (with different structures) to represent the same information. Finally, semantic heterogeneity is caused by divergent interpretations of data according to the different contexts in which such data can be used. Semantic heterogeneity, which is the focus of this work, has been frequently approached with the use of ontologies [4].

As presented in [5][6], ontologies can be used, among other possibilities, as global (or shared) conceptualization for data integration. In this sense, ontologies can promote data interoperability by providing a common semantic background for data interpretation, reducing conceptual ambiguities and inconsistencies, and supporting meaning negotiation. In the last decades, several ontologies have been built for this purpose. In some success cases, they have become reference models reused by a large community, e.g., the *Gene Ontology* proposed by [7] has had a significant impact in the sharing of scientific knowledge about the functions of genes. In other cases, they have failed to establish de facto shareability, and consequently to support data interoperability.

This failure may have many reasons. A relevant one surfaces when new ontologies are developed disregarding previous knowledge resources (i.e., any type of artifact that represents knowledge about a domain, including ontologies and other kinds of reference models and representation schemes). This creates new interoperability problems among existing

ontologies. As a result, reuse has become a common concern in the ontology engineering area [8][9].

Some ontology engineering methodologies describe specific activities to deal with reuse [10][11]. Despite that, many challenges still need to be tackled to promote reuse. The NeOn methodology [10], for example, proposes eight scenarios for building ontologies from the reuse of previous knowledge resources. However, NeOn provides only generic guidelines for the search and selection of reusable knowledge resources. Since no other ontology engineering methodology consulted provides a systematic method for accomplishing these activities, we realize the need to propose an approach to do so in a systematic way.

Even when systematic strategies for searching and selecting reusable knowledge resources are available, we still often have to deal with the integration of different knowledge resources. In the case of environmental data, there is an ample spectrum of subject domains covered (geographical features, spatial coordinates, environmental quality parameters, measurement activities, sampling activities, involved organizations, etc.). Given this broad spectrum, none of the available knowledge resources on their own covers all subject domains needed to integrate such data. Because of this, existing knowledge resources need to be integrated. Thus, we decided to propose the design of a reference ontology for the integration of environmental data, based on the combined reuse of some of these knowledge resources.

It is worth mentioning that the reusable knowledge resources on environmental domain are not always explicitly related, which makes them unsuitable for data integration in their current form. Consequently, some effort is required for their integration. This task will be addressed in this work with the adoption of a common foundational ontology [12]. A foundational ontology provides basic notions of object, relation, property, event, and others. This makes it possible to correlate and, if necessary, adapt the elements of different knowledge resources.

This work is inserted in a project entitled “An eScience Infrastructure for Water Quality Management in the Doce River Basin”, called henceforth Doce River Project for brevity. This project is concerned with the integration of water quality data produced by various sources to assess the impacts of the mining disaster that occurred in the city of Mariana, in Brazil, in 2015, when the Fundão tailings dam broke, contaminating the Doce

River Basin. Thus, the proposed ontology focuses on the integration of water quality data (particularly data from the Doce River Basin).

1.2 Context: The Doce River Project

The Doce River Project originates from the Brazilian environmental disaster that occurred with the rupture of the Fundão tailings dam in the city of Mariana, in the state of Minas Gerais (MG), on November 5th, 2015. This event discharged 55–62 million m³ of iron ore tailings slurry directly into the Doce River Basin, an important basin in the Southeast of Brazil. The mine slurry filled hydrologic networks along 663.2 km of the Doce River through the states of Minas Gerais (MG) and Espírito Santo (ES) before reaching its estuary, in the city of Linhares (ES). As shown in Figure 1, this has caused irreversible environmental damage to hundreds of watercourses and associated ecosystems [13].

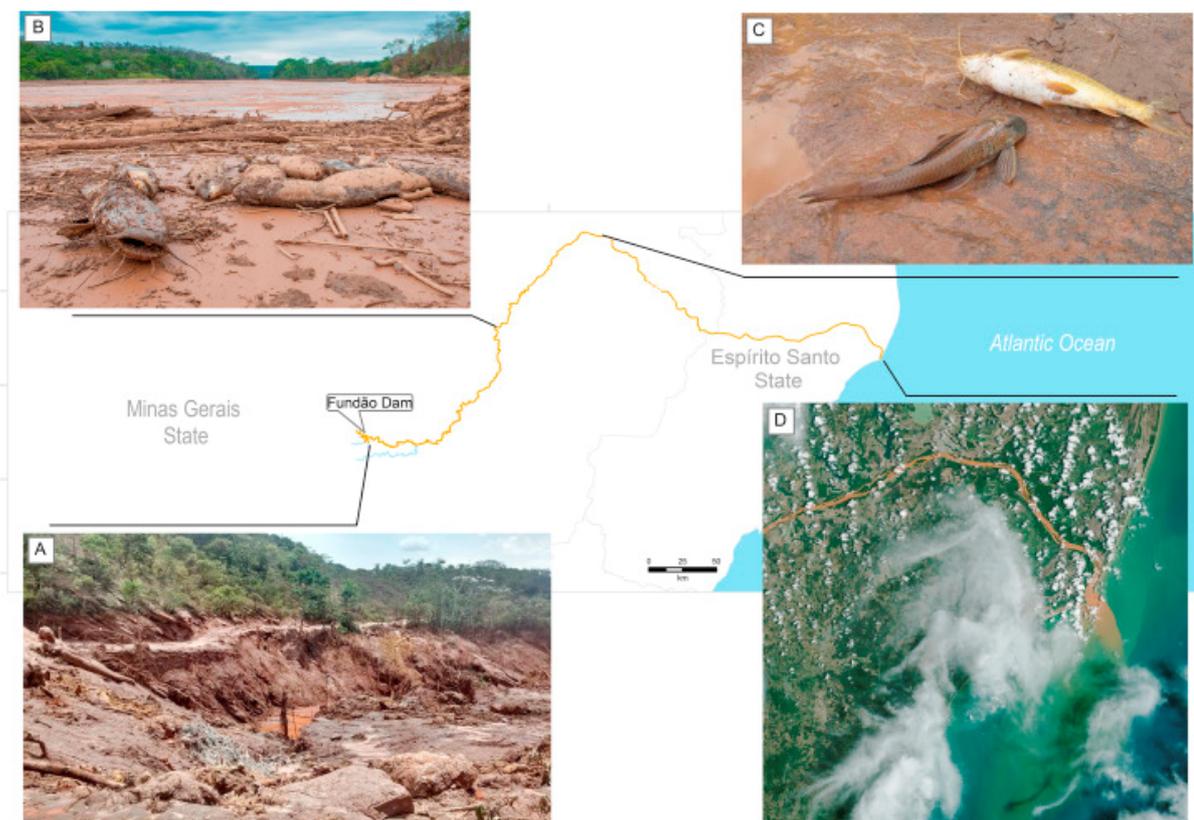


Figure 1 - Impact of the mud wave on the Doce River. (A) the river in the Camargos Municipality few days after the disaster, (B) dead fishes nearby the Doce River Park, (C) dead fishes at Governador Valadares, and (D) the Doce River mouth 25 days after the dam burst [13].

In response to this disaster, autonomous groups of researchers and governmental (e.g., ANA [14], CPRM [15], IBAMA [16], IGAM [17], IEMA [18]) or non-governmental agencies (for example, Renova Foundation [19]) began to take actions to evaluate its consequences,

producing a large volume of data in different knowledge areas (hydrology, geochemistry, biology, among others). In order to support these activities, it is necessary to make these data available, and to support their integrated use. To do this, one has to deal with data heterogeneity problems, and to avoid wrong comparisons when data is obtained by incompatible techniques or when produced for different purposes according to the interest of each data provider.

The research carried out in the Doce River Project aims to produce and analyze water quality data of the Doce River Basin as an attempt to answer questions about the water quality in the basin in general, and, more specifically, concerning the impact of the disaster on the affected environment. To enable the integration of heterogeneous data, the project aims to develop an ontology to provide a shared conceptualization for these data. The ultimate goal is the development of an e-Science infrastructure [20] based on this ontology for the publication of such data according to the principles of FAIR Data [21]. To achieve these goals, the project counts with a team composed of researchers from the areas of Geochemistry, Aquatic Biodiversity and Computer Science.

1.3 Objectives

This work has the main objective of developing a reference ontology to enable the integration of water quality data from the Doce River Basin. Such data are heterogeneous, produced by many sources for different purposes. Thus, the reference ontology has the purpose of serving as a shared conceptualization to solve the semantic heterogeneity caused by divergent interpretations of data according to the different contexts in which they are used.

To avoid the unnecessary proliferation of new ontologies, we have decided to reuse existing knowledge resources on the environmental domain. However, reuse-focused ontology engineering methodologies present very general guidelines for the search and selection of knowledge resources to be reused. Thus, a second objective of this work is to propose an approach to perform these activities in a systematic way.

1.4 Approach

To develop the reference ontology for the integration of water quality data, we chose to follow some guidelines from the NeOn methodology [10]. This is because NeOn focuses on

the reuse of existing knowledge resources. Since NeOn provides only generic guidelines for the search and selection of reusable knowledge resources and no other ontology engineering methodology consulted provides a systematic method for accomplishing these activities, we propose an approach to carry them out systematically. The approach is dubbed CLeAR (Conducting Literature Search for Artifact Reuse). As CLeAR focuses on specific activities in the ontology engineering process, it should be embedded in a comprehensive ontology engineering methodology (such as NeOn).

CLeAR is based on some practices of the Systematic Literature Review (SLR) [22][23]. The search in the scientific literature becomes the basis for the identification of knowledge resources that jointly cover the domain and exhibit properties considered desirable for reuse (proper documentation, community acceptance, among others). In general, CLeAR activities consists of: (i) defining data integration requirements; (ii) finding reusable knowledge resources on the domain of interest; and (iii) selecting some of the identified knowledge resources to be reused in the development of ontology for data integration purposes.

In order to define data integration and, consequently, ontology requirements, CLeAR proposes the use of both top-down and bottom-up analysis. The top-down analysis is performed through the definition of integration questions (IQs) driven by the needs of domain experts. IQs are questions about the domain that can only be answered through the integration of different data sources. The bottom-up analysis is done by studying the elements of the data sources to be integrated. This enables the identification of the domain aspects. Domain aspects are elements of the domain that can be handled in a modular way (e.g., research activities, actors and roles description, and characterization of researched entities). They need to be covered by the reusable knowledge resources.

We have applied CLeAR to the water quality domain. A total of 543 publications were surveyed. The results obtained provide a set of 75 knowledge resources on this domain. This set of knowledge resources make up a knowledge base on the domain to be revisited and reused whenever necessary. This justifies the effort employed in performing the systematic search for a domain for the first time.

Six of the retrieved knowledge resources were selected for reuse in the development of the proposed ontology. However, as they differ from each other and cannot be integrated into

their original format, it was necessary to perform an ontological analysis of them based on a foundational ontology [12]. This analysis reveals the correspondences between the knowledge resources elements and concepts in the foundational ontology. This makes it possible to adjust previous knowledge resources or portions of them for integration into the proposed ontology.

Particularly, to model the water quality domain, we need the general concept of events to deal with research activities (sampling, measurement, etc.); the basic concept of object to represent geographic features (river, lake, etc.), material entities (e.g., water, sediment), devices and procedures used by the research activities, etc.; the concept of agent, to deal with involved people and organizations; the concepts related to qualities, to account for environmental quality parameters and their quantification; and so on. As the Unified Foundational Ontology (UFO) [24][25][26] provides these basic concepts, we have used UFO to analyze the reusable knowledge resources and ground the proposed ontology.

Moreover, we realize that in environmental research there are many general concepts that are applicable across a number of (sub) domains. For example, the concepts related to research activities, spatial location (geographic features and geographic coordinates) and material entities are pervasive notions in environmental research. Thus, they can be represented by means of core ontologies. Core ontologies provide a precise definition of structural knowledge in a specific field that spans across different application domains in this field [27]. They can be reused and extended to incorporate particularities of the domains of interest, that is, for the construction of domain ontologies.

Due to these characteristics and the complexity of the environmental domain, we have decided to modularize the ontology into an ontology network [10]. This facilitates the maintenance and growth of the ontology. The architecture adopted to organize such ontology network, proposed by [28], is shown in Figure 2.



Figure 2 - Ontology Network Architecture proposed by [28].

At the bottom layer, UFO [24][25][26] is used to provide the general ground knowledge for classifying concepts and relations. In the center, core ontologies are used to

represent the general domain knowledge (about research activities, spatial location, etc.), being the basis for the sub (domain) networked ontologies. Finally, (sub) domain ontologies reusing foundational and core ontologies are used to describe the more specific knowledge (about water quality and environmental monitoring).

It is noteworthy that most of the concepts of the networked ontologies were reused from the knowledge resources selected for reuse with the application of CLeAR to the water quality domain. New concepts have been added as needed.

Lastly, we evaluated the proposed ontology network. For that, we have checked whether the elements of the ontology network can support answering each of the integration questions defined during the ontology requirements definition. In addition, we have shown how the elements of the data sources to be integrated correspond to concepts in the ontology network. Figure 3 presents the various activities that were performed in the development of this work.

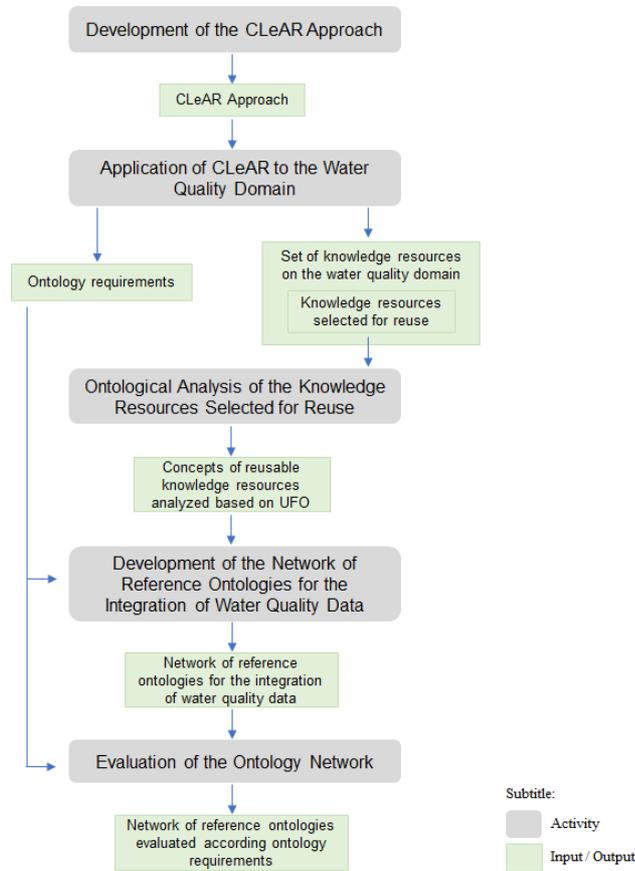


Figure 3 - Activities performed in the development of this work.

1.5 Structure

The remainder of this work is organized as follows.

Chapter 2 presents some background knowledge that supports our investigation on the development of ontology networks with reuse. The chapter addresses ontologies, ontology networks, ontology engineering methodologies (in particular the NeOn methodology), and identifies gaps of these methodologies related to reuse. In addition, it presents an overview of Systematic Literature Review (SLR) practices used in this work as inspiration for searching and selecting existing knowledge resources.

Chapter 3 describes the CLeAR approach. CLeAR is a systematic approach to find and select reusable knowledge resources for building ontologies with the purpose of scientific research data integration. CLeAR adopts some practices of SLR.

Chapter 4 discusses the results of the application of CLeAR to the water quality domain in the context of the Doce River Project. The objective is to find and select existing knowledge resources to be reused in the development of the network of reference ontologies for the integration of water quality data.

Chapter 5 presents an ontological analysis of the knowledge resources selected for reuse based on UFO, focusing on their concepts that are relevant to this work.

Chapter 6 presents and evaluates the network of reference ontologies for the integration of water quality data. It was developed based on the knowledge resources selected for reuse in Chapter 4 and is organized in the layered architecture previously presented.

Chapter 7 discusses final considerations and future work. A summary of the main contributions is provided, the difficulties and limitations are discussed and future research directions are presented.

2 Background

In this chapter, we review some background knowledge that was required for the development of this work. They include ontologies, ontology networks, ontology engineering methodologies, in particular the NeOn methodology [10], gaps of existing methodologies related to reuse, and the Systematic Literature Review (SLR) [22][23].

2.1 Ontologies

The term “ontology” has its origin in Philosophy and refers to both a philosophical discipline (Ontology with a capital “O”) and a domain-independent system of categories that can be used in the conceptualization of domain-specific scientific theories. Since its introduction in Computer and Information Science literature in 1967, ontology has become popular and has been used with different senses by different communities. In information systems, ontology is used in ways that conform to its definitions in philosophy. As a system of categories, an ontology is independent of language. In contrast, in Artificial Intelligence and Semantic Web communities, ontology is, in general, a concrete engineering artifact designed for a specific purpose and represented in a specific language. Languages, formalisms, and tools to create, store and communicate ontologies have proliferated in recent years (e.g., KIF, Ontolingua, UML, OWL) [29].

We have adopted the ontology definition presented by [30] where “*An ontology is a formal, explicit specification of a shared conceptualization*”. In this definition, “conceptualization” refers to a set of relevant concepts and relations used to represent some phenomenon of the real world. “Explicit” means that the type of concepts used, and the constraints on their use are explicitly defined. “Formal” refers to the fact that that the ontology should be machine readable, which excludes natural language. “Shared” reflects the notion that an ontology captures consensual knowledge, that is, it is not private to some individual, but accepted by a group.

In this sense, an ontology can be seen as an engineering artifact defined in terms of classes or types of entities, their properties and relations, along with axioms to establish the admissible combinations of entities in a given domain. In addition, an ontology may define instances of the types considered.

Ontologies can be classified in several ways. One distinguishes ontologies according to their level of abstraction in: (i) foundational (or top-level) ontologies that span across many fields and model very basic and general concepts and relations that make up the world, such as space, time, matter, object, event, action, etc. [12]; (ii) core ontologies that provide a precise definition of structural knowledge in a specific field that spans across different application domains in this field (they are built based on foundational ontologies and provide a refinement to them by adding detailed concepts and relations in their specific field) [27]; and domain ontologies that represent knowledge about a particular domain (they are based on foundational or core ontologies by specializing their concepts) [27]. In this work, ontologies employed cover these various levels of abstraction.

Another classification takes into account a representation's intended use and differentiates ontologies as conceptual models, called reference ontologies, from ontologies as computational artifacts, called operational ontologies [29]. A reference ontology is constructed with the goal of making the best possible description of the domain in reality, representing a model of consensus within a community, regardless of its computational properties. Once users have already agreed on a common conceptualization, operational versions (machine-readable ontologies) of a reference ontology can be implemented. Contrary to reference ontologies, operational ontologies are designed with the focus on guaranteeing desirable computational properties [11]. In this work, we are concerned primarily with the design of reference ontologies.

2.2 Ontology Network

The representation of complex domains through a single ontology leads to the creation of large monolithic ontologies that are difficult to reuse and maintain. In such cases, modularization must be considered as a way of structuring ontologies. This means that the development of a large ontology must be based on the combination of self-contained, independent and reusable knowledge components [31]. An ontology network is essentially a modular ontology, made of components (the individual ontologies) related together via a variety of relationships, such as alignment, modularization, and dependency. A networked ontology, in turn, is an ontology included in such a network, sharing concepts and relations with other ontologies. This representation favors the reuse, maintenance and growth of the model [10].

In [28], it is argued that an ontology network must be equipped with mechanisms that allow it to be gradually improved and expanded. Thus, an ontology network must take into account three main premises: (i) be based on a well-founded grounding for ontology development; (ii) offer mechanisms to easy building and integrating new (sub) domain ontologies; and (iii) promote integration by keeping a consistent semantics for concepts and relations along the ontology network. As shown in Figure 2, a layered architecture is proposed to organize the ontology network. At the bottom layer, a foundational ontology is used to provide the general ground knowledge for classifying concepts and relations. In the center, core ontologies are used to represent the general domain knowledge, being the basis for the sub (domain) networked ontologies. Finally, on top of core and domain ontologies, (sub) domain ontologies are used to describe the more specific knowledge.

There are three different ways to incorporate ontologies to the ontology network, considering the origin of the ontology to be integrated. In the first way, new ontologies are created based on foundational and/or core ontologies, and also taking other existing networked ontologies into account. Besides the extensions made from the foundational/core ontologies, they tend to use the related concepts already defined in the other networked ontologies. This is the best way for increasing the ontology network, since it reduces modeling and integration efforts, by reusing already defined elements [28].

In the second way, new ontologies are developed based on foundational and/or core ontologies, however, independently of the other networked ontologies. Thus, some additional integration effort is still required to adapt the common parts focusing on a shared representation [28].

In the third way, external ontologies, developed without taking the foundational and/or core ontologies as basis, are integrated to the ontology network. In this case, if one has access to modify these ontologies, it is necessary to perform an ontological analysis and reengineering them before the integration. By this process, the ontologies elements are analyzed and adapted to the foundational and/or core ontologies concepts. The knowledge represented by the external ontologies is then preserved, but the representation is adjusted for a better integration into the ontology network. On the other hand, if the ontology cannot be modified, one must to make the necessary links and adaptations only in the ontology network side. In this case, techniques for ontology alignment apply [28].

In this work, the proposed ontology network will be organized in this layered architecture. Existing knowledge resources will be analyzed based on a foundational ontology and adapted, if necessary, to be integrated into the ontology network.

2.3 Ontology Engineering Methodologies

Ontology Engineering is formally defined as “*the set of activities that concern the ontology development process, the ontology life cycle, and the methodologies, tools and languages for building ontologies*” [32]. Ontology engineering methodologies provide guidelines for the development, management and maintenance of ontologies. Such methodologies decompose the ontology engineering process in a number of steps, and recommend activities and tasks to be performed for each one. In addition, they define the roles of the individuals and organizations involved in the ontology engineering process. In general, *domain experts* provide knowledge with respect to the domain to be modeled, *ontology engineers* have expertise in fields such as knowledge representation and development tools, and *users* apply the ontology for a particular purpose [33].

In [32], the authors differentiate three types of activities within an ontology engineering process: management, development and support activities (see Figure 4). The first covers the organizational setting of the overall process. In particular, at pre-development time, a feasibility study examines if an ontology-based application, or the use of an ontology in a given context is the right way to solve the problem at hand. The second type of activities refers to classical activities such as domain analysis, conceptualization and implementation, but also maintenance and use, which are performed at post-development time. Ontology support activities such as knowledge acquisition, evaluation, reuse, and documentation are performed in parallel to the development activities [33].

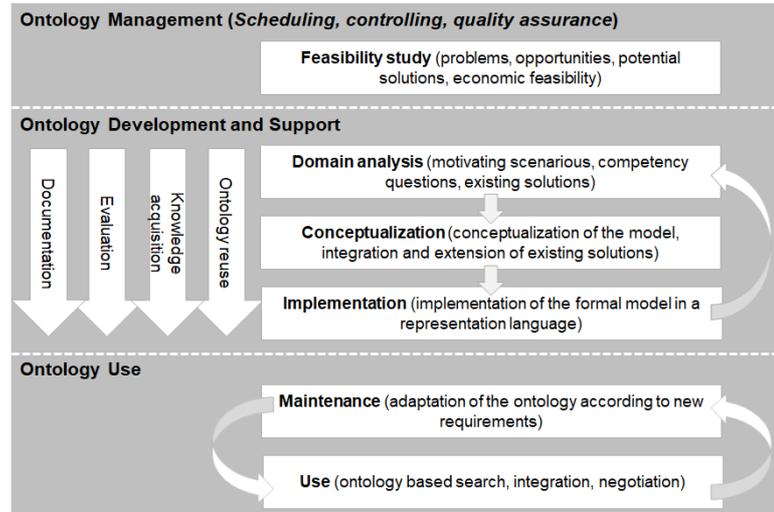


Figure 4 - Main Activities in Ontology Engineering extracted from [33].

A distinction between ontology engineering methodologies takes into account the strategy adopted for building ontologies, that is, building from scratch or building from existing knowledge resources [32]. Examples of methodologies that address building ontologies from scratch can be found in [34]. As examples of methodologies that describe specific activities for addressing reuse, we can cite NeOn [10] and SABiO [11]. These methodologies advocate that the reuse of previous knowledge resources enables speeding up the ontology development process and avoids the proliferation of unnecessarily new models [8][9]. Next, we present the NeOn methodology [10] because in this work we have adopted some of its guidelines for the development of the proposed ontology network.

2.3.1 The NeOn Methodology

The NeOn methodology [10] provides nine possible scenarios for building ontologies and ontology networks. Eight of them are designed to deal with the reuse of existing knowledge resources [10]. In general, the activities proposed by these scenarios are: (i) specification of ontology requirements, (ii) search for reusable knowledge resources, (iii) assessment of candidate knowledge resources, (iv) selection of knowledge resources, (v) adaptation of selected knowledge resources (reengineering, alignment, merging, etc.), (vi) ontology conceptualization, (vii) ontology formalization, (viii) ontology implementation, and (ix) ontology evaluation.

The objective of the activity “specification of ontology requirements” is to output the ontology requirements specification document (ORS) that includes the purpose, the scope, and the implementation language of the ontology network, the target group, and the intended

uses of the ontology network, as well as the set of requirements that the ontology network should fulfill, mainly in the form of competency questions (CQs) [10]. CQs are questions written in natural language that the ontology should be able to answer [35]. This activity is performed by ontology developers (ontology engineers), domain experts and users [10].

After the specification of ontology requirements, it is recommended to carry out a search for candidate knowledge resources to be reused in order to speed up the ontology development process. These knowledge resources can be ontologies, non-ontological resources (e.g., thesauri, glossaries, databases) and ontology design patterns. Ontology developers and domain experts use the terms that have the highest frequency in the ORSD to search for non-ontological resources that cover the desired terminology. This search is performed in highly reliable websites, domain-related sites, and resources within organizations [10].

In the case of the search for ontologies, ontology developers reformulate CQs with vocabulary that may belong to reusable ontologies but that do not explicitly appear in CQs. They identify definitions and axioms that can be potentially reused in the ontology to be developed. The terms whose definition could be reusable from other ontologies are those appearing in the ORSD and the reformulated CQs. Ontology developers search for ontologies that implement these definitions and axioms in general purpose search engines (e.g., Google), Semantic Web search engines (e.g., Swoogle, Watson), and repositories (e.g., the Protégé ontology library, the Open Biological and Biomedical Ontologies). The output of the “search for reusable knowledge resources” is the set of candidate knowledge resources to be reused [10].

In the next activity, “assessment of candidate knowledge resources”, the set of candidate knowledge resources obtained is assessed. Ontology developers must inspect the content and granularity of ontological resources to verify that they meet the needs identified in the ORSD. Ontologies are compared, taking into account a set of criteria (e.g., reuse economic cost, code clarity, and content quality). Non-ontological resources are assessed by means of the following criteria: coverage, precision, quality and consensus about the knowledge and terminology used in the resource, which is a subjective criterion. Based on the assessment performed, ontology developers select the set of knowledge resources that are the most appropriate for the ontology network requirements [10].

The selected knowledge resources often need to be adapted before being reused in the ontology network. Non-ontological resources are analyzed in order to identify its underlying components and re-engineer them by creating ontological representations of the resource at the different levels of abstraction (e.g., conceptual, computational). Ontologies are adapted through reengineering, alignment, merging, and so on [10].

In the “ontology conceptualization”, ontology developers organize and structure knowledge into meaningful conceptual models at the knowledge level. This activity is independent of the way in which the ontology implementation will be carried out. In the “ontology formalization”, the conceptual model is transformed into a formal or semi-computable model according to a knowledge representation paradigm (e.g., description logics, frames, rules, etc.). In the “ontology implementation”, a computational model (implemented in an ontology language such as OWL) is generated [10].

Finally, “ontology evaluation” is defined as the activity of checking the technical quality of an ontology against a frame of reference. NeOn distinguishes two types of ontology evaluations depending on the frame of reference used [10]:

- Ontology verification is the ontology evaluation activity that compares the ontology against the ontology specification document (ontology requirements and competency questions), thus ensuring that the ontology is built correctly (in compliance with the ontology specification).
- Ontology validation is the ontology evaluation activity that compares the meaning of the ontology definitions against the intended model of the world that aims to conceptualize. In this case, the participation of domain experts and ontology users is essential. Besides expert judgment, another relatively easy way to validate an ontology is by means of instantiation.

2.3.2 Reuse-Related Gaps

Reuse is pointed out as a promising approach to ontology engineering [8], since it enables speeding up the ontology development process and avoids the unnecessary proliferation of new models. As stated earlier, some ontology engineering methodologies such as NeOn [10] and SABiO [11] describe specific activities for addressing reuse. However, SABiO presents

only some types of reuse; and NeOn provides only generic guidelines for searching and selecting reusable knowledge resources.

For example, NeOn [10] instructs ontology developers and domain experts to use the terms that have the highest frequency in the ontology requirements specification document (ORSD) to search for non-ontological resources that cover the desired terminology. This search must be performed in highly reliable websites, domain-related sites, and resources within organizations. In this case, NeOn does not show how to perform the search and record the search results.

To search for ontologies, NeOn [10] suggests that ontology developers reformulate CQs with vocabulary that may belong to reusable ontologies but that do not explicitly appear in CQs. In addition, they have to identify definitions and axioms that can be potentially reused in the ontology to be developed. The terms whose definition could be reusable from other ontologies are those appearing in the ORSD and the reformulated CQs. Ontology developers must search for ontologies that implement these definitions and axioms in general purpose search engines. Besides not showing how to search and record the results, NeOn suggests a subjective process, since one has to make assumptions about terms, definitions and axioms that may belong to reusable ontologies.

For the assessment of candidate resources, NeOn [10] guides ontology developers to inspect the content and granularity of ontologies to verify that they meet the needs identified in the ORSD. Ontologies must be compared, taking into account a set of criteria (e.g., reuse economic cost, code clarity, and content quality). Non-ontological resources are assessed by means of the following criteria: coverage, precision, quality and consensus about the knowledge and terminology used in the resource, which is a subjective criterion.

2.4 Systematic Literature Review

As we have discussed in the previous section, there is explicit support for reuse in ontology engineering methodologies such as NeOn. However, NeOn provides only generic guidelines for reusable knowledge resources search and selection activities, and no other ontology engineering methodology consulted provides a systematic method for accomplishing them. This justifies a more systematic approach to perform these activities. We draw inspiration for such approach from the practices of the Systematic Literature Review (SLR) [22][23].

SLR is one of the main mechanisms that support evidence-based research. This research paradigm has been advocated as a good practice for decision-making or troubleshooting in many areas such as Medicine, Economics, and Software Engineering [22][36]. An SLR is a secondary study method based on evaluating and interpreting all available research relevant to a particular research question, topic area, or phenomenon of interest, and then on reporting the used methodology and the obtained results. Although an SLR requires considerable effort to be implemented when compared to ad hoc literature reviews, SLRs are auditable, more trustworthy and rigorous [23][37].

An SLR (following [23]) has three phases: planning the review, conducting the review and reporting the review. In the planning phase, the first step is to identify the need for the review, that is, the reason the review is being carried out. Then, the review protocol is developed. A review protocol specifies the methods that will be used to perform a specific SLR. It must contain: the research questions that the review aims to answer; the strategy to search for primary studies, including search terms, search string, and search engines; the criteria and procedures for selecting studies; the checklist and procedures for assessing the quality of studies; the strategy for extracting data; and the strategy for the synthesis of extracted data. The protocol is refined in the following phases, but must be defined in planning to make it less likely that the results of the literature will be biased and search assumptions explicit.

In the conduction phase, the search is performed and the primary studies are retrieved. Next, the selection criteria are applied to identify the studies that provide direct evidence about the research questions. Then, the quality of the selected studies (related to the extent to which the studies minimize bias and maximize internal and external validity) is evaluated. Finally, some data are extracted from the selected studies and synthesized in tables so that the meta-analysis (i.e., statistical techniques aimed at integrating the results of the primary studies) can be performed. In the reporting phase, the main report with final results is prepared and evaluated to verify if the search need has been met [23].

As a way to enhance the quality of the search, Snowballing can be performed [38]. Snowballing refers to using the reference list of a study or the citations to the study to identify additional studies. Using the references and the citations respectively is referred to as

backward and forward snowballing. The studies obtained from the snowballing are analyzed in the same way that the studies returned directly by the search.

In this work, SLR is useful because we are interested in searching for reusable knowledge resources on a scientific research domain. However, we aim to investigate scientific literature and technical papers to find available knowledge resources in the domain of interest. Thus, the SLR planning, conducting, and reporting activities need to be adapted to accommodate this characteristic. This is the subject of CLeAR as discussed in the next chapter.

3 The CLeAR Approach

CLeAR (Conducting Literature Search for Artifact Reuse) is a systematic approach to find and select reusable knowledge resources (here called structured resources) for building ontologies with the purpose of scientific research data integration. By structured resources we mean those that represent knowledge through the use of formal specification of concepts, relations and properties as ontologies, and also other types of artifacts that capture semantic value for the concerned domain, such as reference models, representation schemas (knowledge base schemas, database schemas), data exchange formats, metadata standards, vocabularies, and thesauri.

As discussed in Chapter 2, the proposed approach adopts some practices of the Systematic Literature Review (SLR) [22][23]. More specifically, publications in a given domain are analyzed as a strategy for finding structured resources available on that domain. This aims to increase the scope of the search and reduce the bias, promoting the identification of structured resources that jointly cover the domain and exhibit properties considered desirable for reuse (proper documentation, community acceptance, among others). As a result, the set of retrieved structured resources make up a knowledge base on the domain to be revisited and reused whenever necessary. This justifies the effort employed in performing the systematic search for a domain for the first time.

CLeAR addresses specific ontology engineering activities. As a consequence, it is designed to be used as a complement to existing ontology engineering methodologies such as NeOn [10] and SaBiO [11].

This chapter is structured as follows. Section 3.1 provides an overview of CLeAR activities. Section 3.2 presents activities related to the definition of data integration requirements. Section 3.3 discusses activities that deal with the search for reusable structured resources. Section 3.4 discusses the activities required to select reusable structured resources. Finally, section 3.5 presents concluding remarks.

3.1 Overview of CLeAR Activities

CLeAR is structured in three cycles as shown in Figure 5. The activities of cycle I aim at defining the data integration requirements and the scope of the ontology to be developed.

These requirements are necessary to perform the activities of the other two cycles. The activities of cycle II aim at systematically identifying structured resources candidates to be reused in the development of the ontology, based on the requirements defined in cycle I. Once identified, the structured resources can be selected to be reused, which is the goal of cycle III. The three cycles are intended to be executed in an iterative fashion. In the same way, the activities of each cycle itself should be visited iteratively. As knowledge about the domain is gathered and requirements are refined, new structured resources are identified and should be considered for reuse. CLeAR activities are detailed in the sequel.

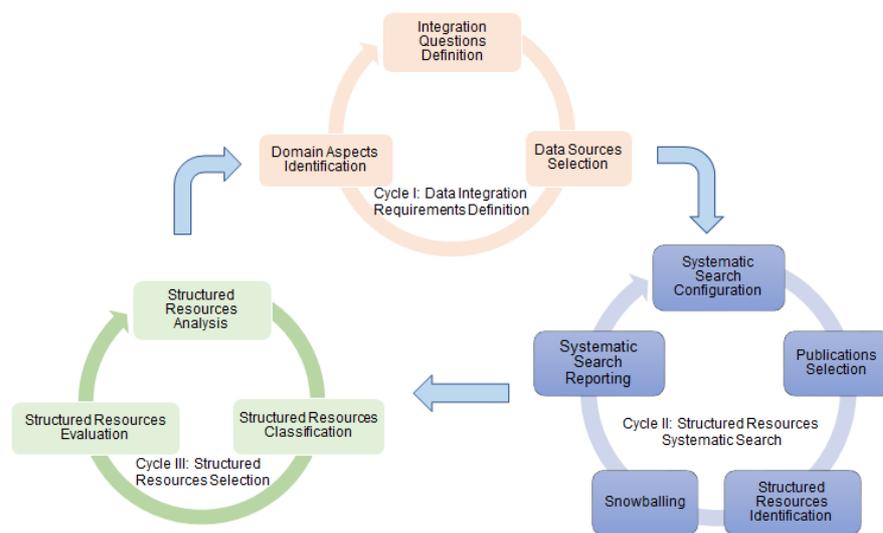


Figure 5 - CLeAR activities.

3.2 Cycle I: Data Integration Requirements Definition

The *Data Integration Requirements Definition cycle (I)* is composed of three activities: (a) *Integration Questions Definition*, (b) *Data Sources Selection* and (c) *Domain Aspects Identification*. In the first activity, a top-down analysis of the integration requirements is made through the definition of integration questions (IQs). IQs are questions about the research domain that can only be answered through the integration of different data sources [3]. In the second activity, the data sources needed to address the IQs are selected by ontology engineers and domain experts. In the third activity of this cycle, a bottom-up analysis of the integration requirements is done by studying the selected data sources. The analysis of data sources, IQs and domain standards combined with the knowledge of domain experts, allows the ontology engineers to identify the domain aspects. Domain aspects are elements of the domain that can be treated in a modular way. They must be enough to represent the universe of discourse.

They are used in cycle II to support the systematic search for structured resources, and in cycle III to guide the selection of structured resources found in cycle II.

3.2.1 Integration Questions Definition

In this activity, a top-down analysis of the integration requirements is made through the definition of integration questions (IQs) driven by the needs of domain experts. IQs are questions about the research domain that can only be answered through the integration of different data sources [3]. That is because the contents of data are different and/or complementary to each other, or because different views of the same content must be contrasted. As IQs are answered from the integration of different data sources, some candidate data sources to be integrated are known to domain experts prior to the application of CLeAR. These data sources serve as input to the definition of IQs. In turn, IQs support the selection of the data sources to be integrated.

As will be seen below, IQs are also used in the definition of the domain aspects. Besides that, in the joint use of CLeAR with ontology engineering methodologies (e.g., [10][11]), IQs are broken down into competency questions. Thus, they are used to define the ontology scope and also for the evaluation of the developed ontology. Since CLeAR is iterative, it allows the refinement of IQs throughout the process, which can be done by adding, grouping, uncoupling and updating actions. Table 1 shows the inputs, outputs and actors of this activity.

Table 1 - Inputs, Outputs and Actors of Integration Questions Definition

Integration Questions Definition	
Inputs	Needs for knowledge about a particular research domain and candidate data sources to be integrated to provide this knowledge
Outputs	Integration questions (IQs)
Actors	Domain Experts

3.2.2 Data Sources Selection

After the definition of IQs, data sources required to answer them are selected by ontology engineers and domain experts. These data sources will be integrated with the support of the ontology to be developed from the reuse of the found structured resources. The selection of data sources can be challenging considering that: (i) data producers may be many

(researchers, government entities, non-profit organizations, industry and laboratories) and sometimes unknown; (ii) data can be difficult to find and obtain due to organizational barriers; and (iii) data can be large, heterogeneous and of varying quality. Table 2 shows the inputs, outputs and actors of this activity.

Table 2 - Inputs, Outputs and Actors of Data Sources Selection

Data Sources Selection	
Inputs	Candidate data sources to be integrated and integration questions (IQs)
Outputs	Data sources to be integrated
Actors	Ontology Engineers and Domain Experts

3.2.3 Domain Aspects Identification

In this activity, a bottom-up analysis of the integration requirements is done by studying the selected data sources. The analysis of data sources, IQs and domain standards combined with the knowledge of domain experts, allows the ontology engineers to identify the domain aspects. Domain aspects are elements of the domain that can be treated in a modular way. They must be enough to represent the universe of discourse. That is, any information about the domain must be part of a domain aspect. They can be related to activities, actors and roles description, characterization of researched entities, and so on.

To define domain aspects, one can use general questions to characterize a scientific research. Examples of these questions are: “How is scientific research done?”, “Where?”, “When?”, “What is researched?”, “Who is the agent or principal?” and “Why is scientific research done?”. Similarly to IQs, domain aspects can be refined continuously by adding, grouping, uncoupling or updating actions. They are used in cycle II to support the systematic search for structured resources, and in cycle III to guide the selection of structured resources found in cycle II.

It is important to note that the analysis of the selected data sources elements provides significant knowledge for the identification of domain aspects. This is because our ultimate goal is to find structured resources to be reused in the development of ontologies for the integration of these data sources. However, as mentioned before, data sources content can be large, heterogeneous and of varying quality. Therefore, care must be taken when analyzing it to identify domain aspects. This involves: correlating different terms used to represent the

same concept; understanding the different granularities used to represent data; and verifying the meaning of the absence of data when not justified. This should be done with the support of the domain experts.

Table 3 shows the inputs, outputs and actors of this activity.

Table 3 - Inputs, Outputs and Actors of Domain Aspects Identification

Domain Aspects Identification	
Inputs	Data sources to be integrated, integration questions (IQs), domain standards, and knowledge of domain experts
Outputs	List of domain aspects
Actors	Ontology Engineers and Domain Experts

3.3 Cycle II: Structured Resources Systematic Search

The *Structured Resources Systematic Search cycle (II)* is mostly inspired in practices of SLR [22][23]. CLeAR, unlike SLR, investigates scientific literature and technical papers to find available structured resources in the domain of interest. Thus, the SLR planning, conducting, and reporting activities were adapted to accommodate this characteristic. In CLeAR, the planning activity is called (a) *Systematic Search Configuration*. The conducting activity is divided into three: (b) *Publications Selection*, (c) *Structured Resources Identification*, and (d) *Snowballing*. The reporting activity is called (e) *Systematic Search Reporting*. They are performed by ontology engineers who are interested in finding structured resources to improve their work.

In *Systematic Search Configuration*, the strategy required to perform the search is defined. Steps such as the specification of the search goals and the definition of inclusion and exclusion criteria are executed. In *Publications Selection*, the systematic search for publications is performed. The returned publications are analyzed and selected by applying the inclusion and exclusion criteria of publications. After the publications selection, the structured resources presented or mentioned by the selected publications are analyzed and selected by applying the inclusion and exclusion criteria of structured resources. This is done in the *Structured Resources Identification* activity. To enhance the quality of the search, the *Snowballing* activity can be performed. The snowballing technique [38] can be applied to both publications and structured resources. As a result of these activities, we have the sets of

identified and selected publications and structured resources. Finally, in *Systematic Search Reporting*, the results of the systematic search are presented and evaluated to verify if the search goals were reached.

3.3.1 Systematic Search Configuration

In *Systematic Search Configuration*, the following steps are executed: specification of the search goals (which concerns ultimately the identification of structured resources in the particular research domain); selection of keywords to compose the search string; elaboration of the search string; selection of search engines; definition of inclusion and exclusion criteria whose purpose is to select only publications and structured resources that meet the search goals; definition of the publications selection procedure; definition of the structured resources identification procedure; and definition of the snowballing procedure.

In CLeAR, the selection of keywords reflects the dual nature of the search goals. Thus, keywords represent not only the domain but also the types of structured resources to be found (ontologies, reference models, database schemas, etc.). In addition, there are two different types of inclusion and exclusion criteria (one for publications, the other for structured resources). The eight steps of this activity are explained below.

Search Goals Specification. In this first step, the search goals are specified to guide systematic search activities. They must be related to the structured resources to be searched.

Keywords Selection. In this step, the terms to compose the search string are selected. Once we are searching for structured resources on a specific domain, we need to define some keywords related to structured resources and others related to the domain. To make reference to structured resources, terms such as “ontology”, “reference model”, “vocabulary”, “taxonomy” and their related terms must be considered. Regarding the domain, keywords that depict the domain itself, the super domain (i.e., a domain more generic than ours) or the domain aspects should be used.

Search String Improvement. The terms obtained in the previous step are organized in a search string. This string should group the keywords into a logical expression (typically using OR and AND operators). In CLeAR, the expression is formed by two main terms connected by AND: the first one selects publications concerned with structured resources and the second one selects domain-specific publications. Each of these main terms is disjunctive in order to

include alternative terms that are used to denote structured resources and to identify the research domain. The search string is tested gradually, including terms subsequently in the disjunctions, in order to test whether they actually increase the search results and should be kept in the string.

Search Engines Selection. After the search string was constructed, the search engines to be used need to be selected. They include digital libraries, specific journals and conference proceedings as recommended by [23]. Checking search engines results against lists of already known primary studies, called here control papers, can be useful for selection of the search engines [23].

Inclusion and Exclusion Criteria Definition. In this step, the criteria to select (inclusion) or discard (exclusion) publications and structured resources obtained by the systematic search are defined. Then, only those that directly reach the search goals are maintained. For publications, a general inclusion criteria recommended by CLeAR is that the publications must present or mention structured resources about the domain or an aspect of it. Other inclusion criteria could be: language, journal, authors, setting, participants or subjects, research design, sampling method and date of publication [23]. For structured resources, an inclusion criteria proposed by CLeAR is that they must address the domain or its aspects. As exclusion criteria, both for publications and structured resources we can check their availability. That is, publications and structured resources whose content is not fully available must be excluded.

Publications Selection Procedure Definition. In this step, the process to be followed for the publications selection is defined. Initially, one must determine the scope of the search, that is, if the string terms will be searched only in title, abstract, or any part of the publications. Second, one must define data to be registered about the publications and the form to be used to record them. Regarding publications data, it is necessary to register: the year, the title, the authors and the source. Additional information may be added.

Structured Resources Identification Procedure Definition. In this step, the process to be followed for the structured resources identification is defined. One must define data to be registered about the structured resources and the form to be used to record them. In relation to the structured resources data, it is necessary to register: the name, the source, the language, the owner, the description, the key concepts, the upper level ontology (applicable only to

ontologies), the resources that reuse the structured resource, the selected publications that present the structured resource, and the selected publications that mention the structured resource. Additional items may be added.

Snowballing Procedure Definition. As a way to enhance the quality of the search, snowballing [38] can be performed. In CLeAR, the snowballing technique has been adapted to be applied to both publications and structured resources. In the case of publications, it can be used in the same way as in the SLR, that is, by checking the reference lists and citations of selected publications. In the case of structured resources, it selects structured resources that are reused by each one analyzed.

Table 4 shows the inputs, outputs and actors of the *Systematic Search Configuration*.

Table 4 - Inputs, Outputs and Actors of Systematic Search Configuration

Systematic Search Configuration	
<i>Search Goals Specification</i>	
Inputs	The motivations for the systematic search
Outputs	The systematic search goals
<i>Keywords Selection</i>	
Inputs	The systematic search goals
Outputs	List of keywords related to structured resources, and list of keywords related to domain
<i>Search String Improvement</i>	
Inputs	List of keywords related to structured resources, and list of keywords related to domain
Outputs	Search string
<i>Search Engines Selection</i>	
Inputs	List of control papers
Outputs	Search engines selected
<i>Inclusion and Exclusion Criteria Definition</i>	
Inputs	The systematic search goals
Outputs	List of publications inclusion criteria, list of publications exclusion criteria, list of structured resources inclusion criteria, and list of structured resources exclusion criteria

<i>Publications Selection Procedure Definition</i>	
Inputs	The systematic search goals
Outputs	Process to be followed for the publications selection, form to record publications data
<i>Structured Resources Identification Procedure Definition</i>	
Inputs	The systematic search goals
Outputs	Process to be followed for the structured resources identification, form to record structured resources data
<i>Snowballing Procedure Definition</i>	
Inputs	The systematic search goals
Outputs	Process to be followed for the snowballing
Actors	Ontology Engineers

3.3.2 Publications Selection

In this activity, the process defined in *Publications Selection Procedure Definition* is performed. The search engines are configured according to the search scope and some inclusion and exclusion criteria, such as the publication language, journal, authors and date of publication. Then, the search is performed. The returned publications data are recorded in the publications form. Publications are analyzed and selected by applying the inclusion and exclusion criteria of publications. Table 5 shows the inputs, outputs and actors of this activity.

Table 5 - Inputs, Outputs and Actors of Publications Selection

Publications Selection	
Inputs	Process to be followed for the publications selection, form to record publications data, list of publications inclusion criteria, and list of publications exclusion criteria
Outputs	Selected publications
Actors	Ontology Engineers

3.3.3 Structured Resources Identification

After the publications selection, the process defined in *Structured Resources Identification Procedure Definition* is performed. The structured resources presented or mentioned by the selected publications are identified. The structured resources data are recorded in the structured resources form. Structured resources are analyzed and selected by applying the

inclusion and exclusion criteria of structured resources. Table 6 shows the inputs, outputs and actors of this activity.

Table 6 - Inputs, Outputs and Actors of Structured Resources Identification

Structured Resources Identification	
Inputs	Process to be followed for the structured resources identification, form to record structured resources data, list of structured resources inclusion criteria, and list of structured resources exclusion criteria
Outputs	Selected structured resources
Actors	Ontology Engineers

3.3.4 Snowballing

In this activity, the process defined in *Snowballing Procedure Definition* is performed. The new publications and structured resources data are recorded on the corresponding forms. New publications and structured resources are analyzed and selected by applying the respective inclusion and exclusion criteria. Table 7 shows the inputs, outputs and actors of this activity.

Table 7 - Inputs, Outputs and Actors of Snowballing

Snowballing	
Inputs	Process to be followed for the snowballing, form to record publications data, form to record structured resources data, list of publications inclusion criteria, list of publications exclusion criteria, list of structured resources inclusion criteria, and list of structured resources exclusion criteria
Outputs	Additional selected publications and structured resources
Actors	Ontology Engineers

3.3.5 Systematic Search Reporting

In this activity, the results of the systematic search are presented and evaluated to verify if the search goals were reached. This is done by analyzing (including graphically) some of the information collected about publications and structured resources. Table 8 shows the inputs, outputs and actors of this activity.

Table 8 - Inputs, Outputs and Actors of Systematic Search Reporting

Systematic Search Reporting	
Inputs	Selected structured resources data
Outputs	Systematic search report
Actors	Ontology Engineers

3.4 Cycle III: Structured Resources Selection

The final *Structured Resources Selection cycle (III)* is composed of three activities: (a) *Structured Resources Analysis*, (b) *Structured Resources Classification* and (c) *Structured Resources Evaluation*. In the first activity, the structured resources identified in cycle II are assessed by verifying domain coverage and key quality attributes for reuse (proper documentation, available representation, community acceptance, among others). This allows the classification of the structured resources in the second activity. Finally, in the third activity, the best classified structured resources are evaluated according to their suitability for the representation of extant data. As a final result, we have the selected structured resources to be reused. In addition, we have a set of relevant structured resources in the research domain, classified according to domain coverage and quality attributes. This set of structured resources can be revisited and reused whenever necessary.

3.4.1 Structured Resources Analysis

In this activity, the structured resources identified in cycle II are assessed by verifying domain coverage and key quality attributes for reuse (proper documentation, available representation, community acceptance, among others).

Domain Coverage Analysis. Domain coverage is analyzed based on the domain aspects. This can be verified in several ways: by checking whether or not a domain aspect is covered by structured resources; indicating the degree of domain aspect coverage by structured resources (not covered, covered, largely covered, and fully covered); among others. The domain coverage provides a relevant criterion for making decisions about structured resources reuse. For example, considering the first option, it is verified that each structured resource covers a subset of the domain aspects set identified in cycle I. Thus, if a domain aspect is covered by only one structured resource, this contributes for deciding to select it for reuse. On the other hand, if the domain aspects covered by a structured resource are a subset of the domain

aspects set covered by another resource, this may indicate that the second is a better choice than the first.

In CLeAR, the domain coverage analysis is performed by means of a matrix as shown in Table 9. Each line of the matrix refers to a structured resource and each column refers to a domain aspect. If a domain aspect is covered by a structured resource, the corresponding cell of the matrix must be checked. The domain aspects are grouped according to the questions that answer to characterize a scientific research. The total of domain aspects covered and the total of domain aspects covered in each group by the structured resources are computed.

Table 9 - Structured Resources Domain Coverage Matrix

Domain Coverage																		
Structured Resource Name	How			Where			When			What			Who			Why		
	Domain Aspect 1	Domain Aspect 2	...	Domain Aspect 1	Domain Aspect 2	...	Domain Aspect 1	Domain Aspect 2	...	Domain Aspect 1	Domain Aspect 2	...	Domain Aspect 1	Domain Aspect 2	...	Domain Aspect 1	Domain Aspect 2	...
SR01	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
SR02	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓						
SR03	✓	✓	✓	✓	✓	✓										✓	✓	✓
SR04							✓	✓	✓	✓	✓	✓	✓	✓	✓			
SR05							✓	✓	✓	✓	✓	✓						
SR06										✓	✓	✓	✓	✓	✓			
...	✓	✓	✓															

Quality Attributes Analysis. The quality analysis supports the choice of the structured resources, since it differentiates resources that have similar domain coverage. Relevant quality attributes for reuse include: reuse economic cost (need to acquire a use license, etc.), understandability effort (e.g., quality of the documentation, code clarity), integration effort (modularization, language used, etc.), and reliability (e.g., development team reputation, popularity) [10]. CLeAR adopts the following quality attributes: proper documentation, available representation, and community acceptance. We have prioritized those attributes as they can be evaluated objectively as discussed in the sequel (other attributes may be added if deemed appropriate).

Proper Documentation: It refers to the availability of documentation to facilitate the understanding of structured resources concepts, relationships and properties and, as consequence, to enable their proper use. We check the availability of glossaries and examples of instantiation. Glossaries explain the meaning intended for the concepts that compose the structured resources. Examples of instantiation allow us to understand what is or is not an instance of concepts.

Available Representation: It is related to the availability of a conceptual (graphical) model and the availability of a computational representation, both of which are desirable. The first one is because it promotes a clear and precise description of domain entities for the purposes of communication, learning and problem-solving. The second one is because it provides a machine-readable implementation version of the structured resource. We have used the language of the structured resources, mapped in cycle II, to help in this analysis.

Community Acceptance: This is about a structured resource being considered a domain standard. This can be verified through metrics that show how well it is recognized and used by the community. To assess how much a structured resource is recognized and reused by the community, we use the number of publications that mention the structured resource and the number of resources that reuse it, respectively. We consider as mentioned or reused the resources that obtained at least 50% of the maximum number of mentions or reuse. This is to disregard little mentioned or reused structured resources.

The quality attribute analysis is performed by means of a matrix as shown in Table 10. Each line of the matrix refers to a structured resource and each column refers to a quality attribute. If a structured resource ranks positively in a quality attribute, the corresponding cell in the matrix must be checked. The quantity of quality attributes in which a structured resource is positively classified is calculated in the “Quality Attributes Score” column. If necessary, different weights can be assigned to the quality attributes to compute the score.

Table 10 - Structured Resources Quality Attributes Matrix

Quality Attributes							
Structured Resource Name	Proper Documentation		Available Representation		Community Acceptance		Quality Attributes Score
	Glossary	Examples	Computational Representation	Conceptual (Graphic) Model	Reused	Mentioned	
SR01	✓						1
SR02	✓	✓					2
SR03	✓	✓	✓				3
SR04				✓	✓	✓	3
SR05	✓	✓	✓	✓			4
SR06		✓	✓	✓	✓	✓	5
...	✓	✓	✓	✓	✓	✓	6

Table 11 shows the inputs, outputs and actors of this activity.

Table 11 - Inputs, Outputs and Actors of Structured Resources Analysis

Structured Resources Analysis	
Inputs	Selected structured resources
Outputs	Structured Resources Domain Coverage Matrix, and Structured Resources Quality Attributes Matrix
Actors	Ontology Engineers

3.4.2 Structured Resources Classification

In this activity, the structured resources are classified in each domain aspects group. Thus, those most appropriate to treat the domain aspects of each group are identified. For this, a final score is computed based on the total of domain aspects covered in each group by the structured resources and their quality attributes score. Initially, these values must be normalized in the [0, 1] interval. Then the arithmetic or weighted average of the normalized values is calculated. The structured resources are classified in each group according to this average. Table 12 shows the inputs, outputs and actors of this activity.

Table 12 - Inputs, Outputs and Actors of Structured Resources Classification

Structured Resources Classification	
Inputs	Structured Resources Domain Coverage Matrix, and Structured Resources Quality Attributes Matrix
Outputs	Structured resources classified in each domain aspects group
Actors	Ontology Engineers

3.4.3 Structured Resources Evaluation

In this activity, the best ranked structured resources in each aspects group are selected and evaluated to verify their suitability for the representation of different domain data. This evaluation is performed trying to annotate each element of the data sources selected in cycle I with the concepts (classes), properties and instances made available by each structured resource. As the structured resources are evaluated, they are selected or discarded. If discarded (because they do not properly represent the elements of the target aspects group), the next resources in the classification should be evaluated.

At the end of this activity, we have a set of complementary structured resources to be reused. In addition, we have a set of relevant structured resources in the research domain, classified according to domain coverage and quality attributes. This set of structured resources can be revisited and reused whenever necessary. Table 13 shows the inputs, outputs and actors of this activity.

Table 13 - Inputs, Outputs and Actors of Structured Resources Evaluation

Structured Resources Evaluation	
Inputs	Structured resources classified in each domain aspects group
Outputs	Set of complementary structured resources to be reused
Actors	Ontology Engineers

3.5 Concluding Remarks

In this chapter, we have presented the CLeAR approach. As main advantages of CLeAR, we can cite: (i) its alignment to the needs of ontology building for the purpose of scientific research data integration, since the scope of the ontology is derived from IQs and data to be integrated; (ii) the identification of reusable structured resources in a particular domain through the use of systematic methods to search and select them (this is the main difference

between CLeAR and NeOn); (iii) the evaluation of objective quality attributes (this is also a difference between CLeAR and NeOn, since NeOn adopts some subjective quality attributes); (iv) and the possibility of using it with existing ontology engineering methodologies to support the search and selection for reusable structured resources.

As a disadvantage of CLeAR we point out the effort required for its application to a domain in the first iteration. However, once applied to a particular domain, CLeAR provides a set of evaluated and classified structured resources that can be revisited and reused whenever new needs about such domain arise. We argue that this result justifies the effort employed. In the next chapter, we present the application of CLeAR to the water quality domain.

4 Applying CLeAR to the Water Quality Domain

In this chapter, we apply the CLeAR approach to the water quality domain in the context of the Doce River Project. The objective is to find structured resources to be reused in the development of the network of reference ontologies for the integration of water quality data.

This chapter is structured as follows. Section 4.1 shows the application of the cycle I of CLeAR to the water quality domain. Section 4.2 discusses the application of the cycle II of CLeAR to the water quality domain. Section 4.3 presents the application of the cycle III of CLeAR to the water quality domain. Section 4.4 discusses related work. Finally, section 4.5 presents concluding remarks.

4.1 Definition of the Water Quality Data Integration Requirements

In this section, we present the application of the cycle I of CLeAR to the water quality domain. A key aspect of this cycle is the participation of domain experts, which are knowledgeable of data semantics and which face themselves integration questions in their research activities. In the Doce River Project, they are researchers in the areas of Geochemistry and Aquatic Biodiversity.

4.1.1 Integration Questions for the Water Quality Domain

A non-exhaustive list of IQs defined by domain experts is shown in Table 14. As one can observe, these questions are related to the assessment of water quality at monitoring points along the Doce River and its tributaries. They concern not only the impacts of the disaster but also water quality in general. These questions are answered by analyzing the measurements of the physical, chemical and biological properties of the water and sediment samples and the ecotoxicological essays carried out by different Brazilian organizations.

Table 14 - Integration Questions

Identifier	Integration Question
IQ01	Which monitoring points have appropriate bathing conditions according to the analysis of thermotolerant coliforms?
IQ02	What is the relation between upstream sewage treatment and concentration of thermotolerant coliforms?
IQ03	Which parameters present concentrations above the thresholds established in the applicable legislation for freshwater (357/2005 CONAMA Resolution class 1)?
IQ04	What is the Water Quality Index (WQI) at each monitored point?
IQ05	What is the relation between meteorological and seasonal conditions and water quality?
IQ06	What is the relation between river flow and water quality?
IQ07	What is the BOD (Biochemical Oxygen Demand) / COD (Chemical Oxygen Demand) ratio at the monitoring points?
IQ08	Was there metal contamination at the collection sites prior to the incident?
IQ09	Is there contamination by metals in samples collected after the incident? How much of this contamination is past tense?
IQ10	Do the levels of metals found exceed the values proposed by the legislation?
IQ11	Do sediment metal levels exceed thresholds adopted by environmental agencies?
IQ12	Do the collected water samples present toxicity?
IQ13	What types of toxicity of the water samples?
IQ14	Is toxicity related to contamination levels?

4.1.2 Data Sources to be integrated

The data sources needed to address the IQs are provided by various Brazilian governmental and non-governmental organizations. Among the governmental ones, there are those that cover the national territory and those that cover the states of Minas Gerais and Espírito Santo, bathed by the Doce River and impacted by the disaster. The national governmental organizations selected are: the National Water Agency (ANA) [14], the Geological Survey of Brazil (CPRM) [15] and the Brazilian Institute for the Environment and Renewable Natural Resources (IBAMA) [16]. ANA is the regulatory agency dedicated to enforcing the objectives and guidelines of the Brazilian Water Law. It coordinates the National Hydrometeorological Network that captures, with the support of states and other partners, information such as level, flow and sediment of the rivers or amount of rainfall. CPRM is the official depository of data and information on geology, mineral resources and water resources of the Brazilian territory. IBAMA is an institute, linked to the Ministry of the Environment, which performs actions of national environmental policies, regarding environmental licensing, environmental quality control, authorization of natural resources usage and environmental monitoring and control.

The state-level governmental organizations selected are: the Water Management Institute of Minas Gerais (IGAM) [17] and the Institute of Environment and Water Resources of Espírito Santo (IEMA) [18]. IGAM is an institute linked to the Secretariat of Environment and Sustainable Development of the State of Minas Gerais, whose functions are to plan and promote actions aimed at preserving the quantity and quality of the state's water resources. IEMA is an institute linked to the Secretariat of the Environment and Water Resources of Espírito Santo, with technical, financial and administrative autonomy. Its purpose is to plan, coordinate, execute, supervise and control the activities of the environment, state water resources and natural resources, whose management has been delegated by the union to the state.

The non-governmental organization selected is Renova Foundation [19], that is the entity responsible for the mobilization to repair damages caused by the rupture of the Fundão dam, in Mariana (MG). It is a non-profit organization, which is a result of a legal commitment called a Transaction Term and Adjustment of Conduct (TTAC). It defines the scope of action of the Renova Foundation, which includes 42 programs that unfold in the many projects that are being implemented in the 670 kilometers of impacted area along the Doce River and its tributaries.

4.1.3 Water Quality Domain Aspects

From the IQs presented in Table 14, it is possible to extract many domain aspects that answer the general questions used to characterize a scientific research. Some of them are: *water sampling, water quality analysis, water quality measurement and water quality monitoring (How)*; *water quality properties (parameters) and meteorological aspects (What)*; *location (Where)*; and *normative element (Why)*. For example, the *normative element* domain aspect, which defines water quality and motivates water sampling, water quality analysis, etc., was obtained from IQ03 and IQ11. IQ03 mentions the applicable legislation for freshwater and IQ11 mentions the metal levels thresholds adopted by environmental agencies.

Table 15 was extracted from the *Weekly Water Quality Bulletin (04-Feb-2019)* obtained at the Renova Foundation website [19]. For each element of this table, we have identified a domain aspect: *provenance* (Renova Foundation); *geographical entities* (water courses); *chemical, physical and biological properties of water* (presence of cyanobacteria, electric conductivity, dissolved oxygen and pH); *meteorological aspects* (rain of the period);

units of measurement ($\mu\text{g/L}$, $\mu\text{S/cm}$, mg/L and mm); *sensors used* (telemetric stations); *reference to norms* (357/2005 CONAMA Resolution [39]) and *compliance*.

Table 15 - Fragment of a Table from the Renova Foundation Weekly Water Quality Bulletin (04-Feb-2019)

Automatic station results: The minimum, average and maximum results for the period evaluated in the week of 28-Jan-2019 to 03-Feb-2019 are presented for the parameters: cyanobacteria, electrical conductivity, dissolved oxygen, pH, and accumulated rain in this period.														
Analyzed Parameters														
Telemetric Stations	Water Course	Cyanobacteria ($\mu\text{g/L}$)			Electric Conductivity ($\mu\text{S/cm}$)			Dissolved Oxygen (mg/L)			pH			Rain of the period (mm)
		Min	Avg	Max	Min	Avg	Max	Min	Avg	Max	Min	Avg	Max	Acc
RCA 02	Carmo River	0.0	0.1	0.4	65.6	69.5	73.7	6.7	7.5	8.6	7.2	7.6	8.4	0.0
RDO 01 ¹	Doce River	0.0	0.2	0.4	F	F	F	7.9	8.6	9.7	7.5	7.8	8.5	15.2
RDO 02		NA	NA	NA	59.3	60.9	62.7	7.5	7.8	8.0	7.4	7.5	7.7	NA
RDO 03		0.0	0.1	0.2	58.3	60.1	62.2	6.8	7.2	7.6	7.3	7.5	7.7	0.0
RDO 04		0.2	0.4	0.7	58.6	60.5	61.7	6.9	7.5	8.3	7.6	8.0	8.6	0.0
RDO 05		0.2	0.5	1.8	79.5	99.7	115.8	7.5	7.9	8.2	7.2	7.3	7.5	0.0
RDO 08 ²		0.1	0.2	0.4	78.2	80.6	82.2	5.9	6.7	7.7	7.3	7.6	8.2	0.0
RDO 12		0.0	0.1	0.3	66.9	68.2	69.4	6.7	7.2	7.9	7.3	7.5	8.0	0.0
RDO 16 ³		0.0	0.1	0.5	0.3	108.4	145.9	5.5	6.6	8.4	4.9	7.2	7.8	0.2

Subtitle:
 NA - Not applicable. There is no parameter measurement at the point.
 F - Failure to measure and / or transmit data.
Bold values - results above the limit of the classification class of the 357/2005 CONAMA Resolution for water class II (100 NTU).
 Comments:
¹ RDO 01 - Failed to measure conductivity. The probe is without weekly preventive maintenance due to access prevented by the owner of the property.
² RDO 08 - The cyanobacteria, conductivity, dissolved oxygen and pH parameters were absent from results from 28-Jan-2019 until 29-Jan-2019 at 16:00, due to the of the transmission cable.
³ RDO 16 - The conductivity sensors presented failures due to sensor problems. They were replaced on 02-Feb-2019.

Table 16 presents an analysis of data source elements in two of the data sources we considered (IBAMA-IEMA and IGAM). For each data source element (usually a column name in tabular data provided by a data source), we have identified a domain aspect. Domain aspects group elements that deal with related concepts. The identified domain aspects are: *provenance* (IBAMA-IEMA or IGAM); *geographic coordinates* (altitude, latitude, etc.); *geographical entities* (hydrographic basin, sub basin, water course, among others); *location* (e.g., site, county, station); *temporal references* (date, year, etc.); *sampling*, which encompasses other aspects such as *sampling method*, inferred from the concept of sample type, and *material entity*, inferred from the concept of sample point category; *measurement*, which contain more specific aspects such as *chemical, physical and biological properties* (e.g. alkalinity of bicarbonates), *units of measurement* (mgCaCO_3/L) and *measurement agent* (data source); as well as *normative elements* (framing class of water course). Note that different data sources cover the same domain aspect with different representation schemes.

Table 16 - Concepts of Water Quality used by Brazilian Organizations

Data Source	Data Source Element	Data Examples	Domain Aspect
IBAMA-IEMA	Site	MG Tributaries	Location
	Sample Point Short Name	AFL-06	Location
	Sample Point Long Name	Piranga MG - Upstream	Location
	Sample Point Category	Lotic fresh water, Lotic brakish water	Material Entity
	Lat	-20.383574	Geographic Coordinates
	Long	-42.902283	Geographic Coordinates
	X	718948	Geographic Coordinates
	Y	7744747	Geographic Coordinates
	Z		Geographic Coordinates
	Projection	UTM23S	Geographic Coordinates
	Datum	SIRGAS2000	Geographic Coordinates
	Date	10-Mar-2016 11:00	Temporal References
	Sample Ref	62277-2016	Sampling
	Lab Ref	62277-2016	Sampling
	Data Source	Merieux	Agent
	Sample Type	Superficial	Sampling
Alkalinity of bicarbonates (mgCaCO3/L)	30.6	Measurement	
IGAM	Hydrographic Basin	Doce River	Geographic Entity
	Sub Basin	Piranga River	Geographic Entity
	UPGRH	DO1 - Piranga River	Geographic Entity
	County	PIRANGA (MG)	Location
	Water Course	Piranga River	Geographic Entity
	Description	Piranga River in the city of Piranga	Location
	Framing Class of Water Course	Class 2	Normative Elements
	Station	RD001	Location
	Altitude	610	Geographic Coordinates
	Latitude (Decimal Degrees)	-20.69	Geographic Coordinates
	Latitude (Degrees Minutes Seconds)	-20° 41' 18.661"	Geographic Coordinates
	Longitude (Decimal Degrees)	-43.3	Geographic Coordinates
	Longitude (Degrees Minutes Seconds)	-43° 18' 8.42"	Geographic Coordinates
	Year	2017	Temporal References
	Sampling Date	02-Jul-2017	Temporal References
	Sampling Time	09:15:00	Temporal References
Alkalinity of bicarbonates	18.8	Measurement	

The analysis of the IQs, the domain standards (e.g., [40]) and the selected data sources elements resulted in the following list of the water quality domain aspects: *research activity, sampling, preparation, measurement, analysis, monitoring, sampling method, preparation method, measurement method, analysis method and monitoring method (How); location, geographic coordinates and geographic entity (Where); material entity, abiotic entity, biotic entity, properties, chemical property, physical property, biological property, unit of measurement and meteorological aspects (What); temporal references (When); agent, sensor and provenance (Who); normative elements (Why)*. These aspects together establish the required coverage of the ontology network to be developed.

4.2 Systematic Search for Structured Resources on the Water Quality Domain

Next, we present the application of the cycle II of CLeAR to the water quality domain. It consists in the systematic search for structured resources on this domain.

4.2.1 Configuring the Systematic Search

The following search goal was formulated for the water quality domain:

Find structured resources candidates to be reused in the development of ontologies for data integration in the water quality domain. Identify the structured resources, the language in which they are represented, the location where they are available, the key concepts addressed by them and the resource owner.

Among the keywords related to structured resources we have used “ontology” and “vocabulary” related terms so that publications containing structured vocabularies and taxonomies were also identified (see Table 17 for alternative terms). With respect to the terms related to domain, besides “water quality” itself and its alternative terms, the super domain “environmental quality” was included to make it possible to carry out a wider search (see Table 18).

Table 17 - Keywords related to Structured Resources

Keyword	Related terms (alternative terms)
Ontology	reference model, knowledge base, schema
Vocabulary	taxonomy, thesaurus

Table 18 - Keywords related to Research Domain

Keyword	Related terms (alternative terms)
water quality	water resource, water evaluation, water analysis, water monitoring, water assessment
environmental quality	environmental resource, environmental evaluation, environmental analysis, environmental monitoring, environmental assessment, environment quality, environment resource, environment evaluation, environment analysis, environment monitoring, environment assessment

The final string obtained is presented below:

(ontology OR vocabulary OR "reference model" OR "knowledge base" OR schema OR taxonomy OR thesaurus)

AND

("water quality" OR "water resource" OR "environmental quality" OR "water evaluation" OR "water analysis" OR "water monitoring" OR "water assessment" OR "environmental resource" OR "environmental evaluation" OR "environmental analysis" OR "environmental monitoring" OR "environmental assessment" OR "environment quality" OR "environment resource" OR "environment evaluation" OR "environment analysis" OR "environment monitoring" OR "environment assessment")

The control papers used to aid in the selection of the search engines are listed in Table 19. They were chosen based on a non-systematic search (see [41]), in which it was possible to find publications that propose structured resources suited for the representation of the water quality domain. We selected Google Scholar as the search engine for our systematic search because Google Scholar retrieves technical works in the domain of interest, presented at domain-specific conferences, as well as scientific papers. Unlike other digital libraries (Engineering Village, Scopus and IEEE Explore), the Google Scholar search retrieves all three control papers.

Table 19 - Control Papers

Identifier	Title	Authors	Year
CP01	An Ontology Framework for Water Quality Management	Lule Ahmedi, Edmond Jajaga, Figene Ahmedi	2013
CP02	A Harmonized Vocabulary for Water Quality	Simon J. D. Cox, Bruce A. Simons, Jonathan Yu	2014
CP03	Defining a Water Quality Vocabulary Using QUDT and ChEBI	Bruce A. Simons, Jonathan Yu, Simon J. D. Cox	2013

The publications inclusion and exclusion criteria are shown in Table 20 and the structured resources inclusion and exclusion criteria are shown in Table 21. PIC01 is directly related to the search goal; PIC02 is used to select only publications globally recognized; and PEC01 is used to discard unavailable publications. SRIC01 is used to select only structured resources that address the water quality domain; SREC01 is used to discard structured resources that are also unavailable (because they have been discontinued or because they have not been made available).

Table 20 - Publications Inclusion and Exclusion Criteria

Identifier	Publications Inclusion Criteria
PIC01	The publication presents or mentions structured resources about the water quality domain or its aspects.
PIC02	The publication is written in English.
Identifier	Publications Exclusion Criteria
PEC01	The publication is not available.

Table 21 - Structured Resources Inclusion and Exclusion Criteria

Identifier	Structured Resources Inclusion Criteria
SRIC01	The structured resource addresses the water quality domain or its aspects.
Identifier	Structured Resources Exclusion Criteria
SREC01	The structured resource is not available.

To broaden the scope of the search, it was decided to apply snowballing on the reference lists and citations of the selected publications and on the structured resources reused by those selected.

4.2.2 Selecting Publications

In relation to the search scope, we decided to look for the keywords in the paper title for pragmatic reasons. In this case, we note that even while searching the title, the relevant publications were returned. One way to verify that relevant publications have not been left out is to check if the systematic search returns publications found by previously non-systematic searches. We verify that the publications found by the non-systematic search presented in [41], which propose structured resources suited for the representation of the water quality domain, were returned by the systematic search. Thus, the search scope was configured in the Google Scholar. Besides that, the option to search only publications written in English was checked in the Google Scholar to meet the inclusion criteria PIC02. The systematic search was performed on the June 21th, 2019. The publications returned were analyzed and selected by applying PIC01 and PEC01. In total, 64 publications were obtained. After applying the inclusion and exclusion criteria, 18 were selected. Publication data can be found in the “Publications Selection” table of the dataset [42] provided with this work.

4.2.3 Identifying Structured Resources

The structured resources extracted from selected publications were analyzed and selected by applying SRIC01 and SREC01. In total, 57 structured resources were obtained. After applying the inclusion and exclusion criteria, 44 were selected. Structured resource data can be found in the “Structured Resources Identification” table of the dataset [42].

4.2.4 Applying Snowballing

The application of snowballing on the reference lists and citations of the selected publications resulted in 479 new publications. After applying the publications inclusion and exclusion criteria to them, 67 were selected. For better organization, new publications were listed in the new tables “Reference Lists Selection” and “Citations Selection” (with the same structure as the “Publications Selection” table) of the dataset [42].

The analysis of the new publications resulted in 34 new structured resources. After applying the structured resources inclusion and exclusion criteria to them, 25 were selected. In addition, the application of snowballing on the resources reused by the 60 selected structured resources resulted in 22 new structured resources. After applying the inclusion and exclusion criteria to them, 6 were selected. All structured resources were identified in “Structured Resources Identification” table of the dataset [42].

At the end of the systematic search, 85 publications were selected from a total of 543 analyzed publications. Also, 75 structured resources were selected as candidates for reuse from a total of 113 identified structured resources.

4.2.5 Reporting the Results of the Systematic Search

After conducting the systematic search, its results must be reported and evaluated to verify if the search goals were reached. As previously discussed, the systematic search returned a total of 543 publications, of which 85 (15.7%) were selected for presenting or mentioned structured resources about the water quality domain or part of it. Among the discarded publications (458 publications), 346 publications (75.5%) did not meet inclusion criteria PIC01, 15 (3.3%) did not meet inclusion criteria PIC02 and 97 publications (21.2%) met exclusion criteria PEC01. This means that most publications were discarded because they did not present or mention a structured resource on the domain of interest, that is, they did not meet the systematic search goal.

Regarding the structured resources, a total of 113 structured resources were obtained (counting those extracted from publications and those reused by other resources). Among them, 75 were selected as candidates for reuse and 38 were discarded. Among the 38 structured resources discarded, 20 (52.6%) did not meet inclusion criteria SRIC01 and 18 (47.4%) met exclusion criteria SREC01. Several links provided by publications were broken. In some cases, it was possible to find them elsewhere, but in cases in which it was not possible, structured resources were excluded according to SREC01.

With respect to data extracted about the selected structured resources, we analyze the language, the number of publications that mention these resources (not included the papers that present them) and the number of resources that reuse them. This is useful in evaluating the quality attributes of the structured resources performed in cycle III. The key concepts

treated by the structured resources are also used in cycle III to verify the coverage of the domain by each of them.

Regarding the language, we have found certain convergence. OWL language is used by 38.9% of the structured resources found while schemas written in RDF and XML have reached 22.2%. Only 8.3% use UML, 6.5% use HTML (structured links), and 24.1% use other languages. For this analysis (see graph of Figure 6), resources have been counted more than once according to the number of languages in which they are made available. The language is used to verify the quality attributes related to the representation level of each structured resource in cycle III.

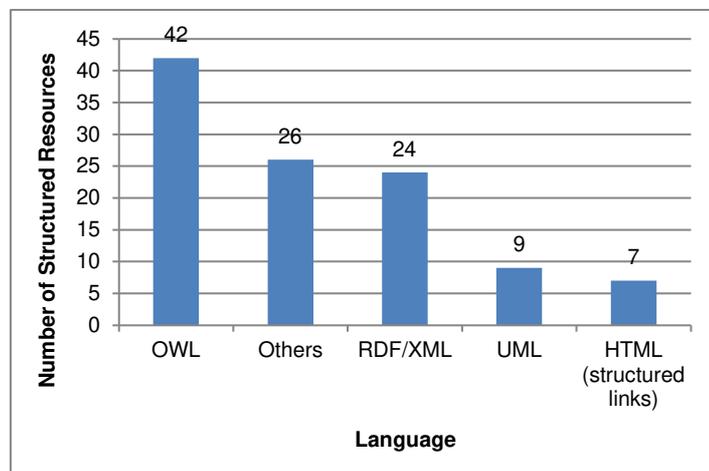


Figure 6 - Language used by the structured resources.

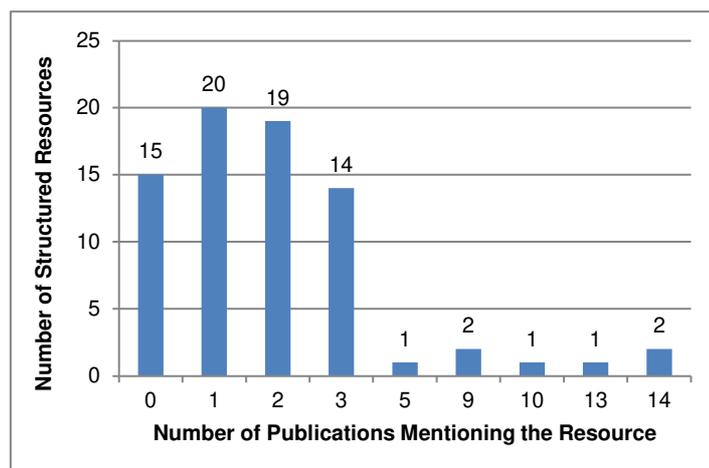


Figure 7 - Popularity of structured resources according to the number of identified publications that mention them.

The number of publications that mention a structured resource can be used to measure how well it is recognized by the community in cycle (III). As shown in the graph of Figure 7,

two structured resources (SSN and SWEET) are mentioned by fourteen publications; one structured resource (O&M) is mentioned by thirteen publications; one resource (ChEBI) is mentioned by ten publications; two resources (OWL-Time and QUDT) by nine publications; and one resource (WaterML) by five publications. 18.7% of the resources are mentioned by three publications; 25.3% of the resources are mentioned by two publications; and 26.7% of the resources by one publication. 20.0% of the structured resources were identified only from the publication that presents them or from the resources that reuse them (they are not mentioned by other publications).

The number of resources that reuse a structured resource represents how much it is used by the community in cycle (III). Regarding the number of resources that reuse a structured resource, the graph of Figure 8 shows that one structured resource (O&M) is reused by twelve resources; one structured resource (GML) is reused by eight resources; one structured resource (SSN) is reused by seven resources; one structured resource (ISO/TC 211) is reused by six resources; one structured resource (OWL-Time) is reused by five resources; and one structured resource (SWEET) is reused by four resources. 2.7% of the structured resources are reused by three resources; 8.0% are reused by two resources; 34.6% are reused by one resource; and 46.7% are not reused by any of the other selected resources.

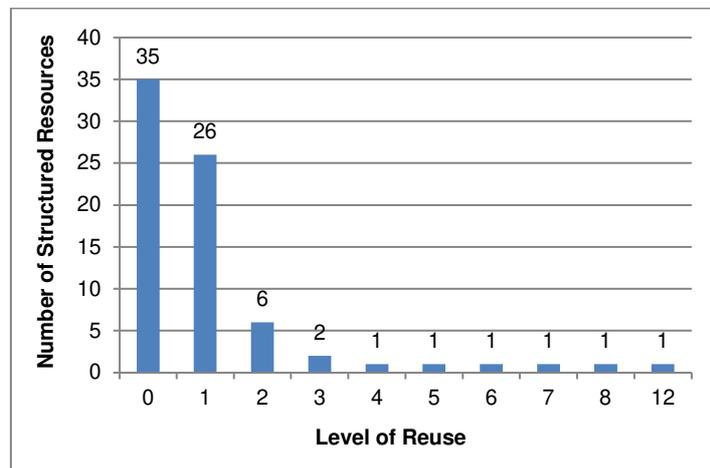


Figure 8 - Level of reuse of structured resources according to the number of structured resources that adopt them.

In relation to the last two graphs, we verify that the structured resources were mentioned or reused by groups different from those that created them. In addition, we disregard the publications that present the structured resources in the analysis performed in

contained in one or two groups. Thus, they tend to be more specific. As examples, we can mention FTT, GeoNames, TGN, USBGN, NGA/GNS, GeoSPARQL and GAZ (Where); OWL-Time and SWRL Temporal (When); and QUDT, OM, QU Rec 20, QU, UCUM, QUDV, NCBITaxon, UO, MDO and ChEBI (What). We do not identify structured resources that cover only domain aspects of How, Who or Why groups.

Table 23 shows the quality attributes analysis for the selected structured resources. The ordering used for Table 22 was maintained to facilitate the identification of the structured resources and the comparison of the two tables. This analysis was recorded in the “Structured Resources Selection” table of the dataset [42].

Table 23 - Quality Attributes for the Structured Resources on the Water Quality Domain

Structured Resource Name	Quality Attributes						Quality Attributes Score
	Proper Documentation		Available Representation		Community Acceptance		
	Glossary	Examples	Computational Representation	Conceptual (Graphic) Model	Reused	Mentioned	
USGS Thesaurus			✓				1
INSPIRE	✓	✓	✓	✓			4
SWEET			✓	✓		✓	3
GEMET		✓	✓				2
ISO/TC 211	✓	✓		✓	✓		4
UsgsHydroML	✓	✓	✓	✓			4
Darwin Core	✓	✓	✓				3
Upper Cyc	✓		✓	✓			3
SUMO			✓	✓			2
InAWaterSense		✓	✓				2
WDTF	✓	✓	✓				3
EML	✓	✓	✓				3
MEMOn	✓		✓				2
GeoSciML	✓	✓	✓	✓			4
EnvO	✓	✓	✓				3
GCMD	✓	✓	✓				3
WaterML	✓	✓	✓	✓			4
ODM	✓	✓	✓	✓			4
O&M	✓	✓	✓	✓	✓	✓	6
CCO			✓				1
EIA			✓				1
EAO	✓			✓			2
WQOP	✓		✓				2
OM-Heavy			✓	✓			2
Wavellite		✓	✓				2
WaWO			✓				1
SAM-Lite	✓	✓	✓	✓			4
WQO			✓				1
WaWO+			✓				1
SERONTO			✓				1
BCO		✓	✓				2
new SSN	✓	✓	✓	✓			4
SensorML	✓	✓	✓	✓			4
GML	✓	✓	✓		✓		4
PEIA		✓	✓				2

ECS	✓			✓				2
OBOE			✓	✓				2
OM-Lite	✓	✓	✓	✓				4
EABS			✓	✓				2
Glossary BAP	✓		✓					2
VSTO			✓	✓				2
SemSOS			✓					1
SEGO	✓	✓	✓	✓				4
Uberon	✓	✓	✓	✓				4
WMO	✓	✓	✓	✓				4
SSN	✓	✓	✓	✓	✓	✓		6
PROV-O	✓	✓	✓	✓				4
WSSN				✓				1
QUDT	✓	✓	✓	✓		✓		5
OM	✓	✓	✓	✓				4
QU Rec 20			✓					1
CF			✓					1
Irstea Hydro			✓	✓				2
MMI			✓	✓				2
WGS84		✓	✓					2
FTT			✓					1
GeoNames	✓	✓	✓					3
TGN	✓	✓	✓					3
USBGN			✓					1
NGA/GNS			✓					1
GeoSPARQL	✓	✓	✓					3
QU	✓		✓					2
UCUM	✓	✓	✓					3
QUDV	✓	✓		✓				3
GAZ			✓					1
NCBITaxon			✓					1
QB	✓	✓	✓	✓				4
EngMath	✓		✓					2
MUO			✓					1
OWL-Time	✓	✓	✓	✓		✓		5
UO			✓					1
SWRL Temporal		✓	✓					2
MDO	✓	✓						2
ChEBI	✓	✓	✓			✓		4
DAML-Time	✓		✓					2

From Table 23, it can be verified that only two structured resources (O&M and SSN) rank positively in all 6 quality attributes; two structured resources (QUDT and OWL-Time) in 5 quality attributes; 24.0% of the structured resources in 4 quality attributes; 16.0% in 3 quality attributes; 30.7% in 2 quality attributes; and 24.0% in 1 quality attribute. 45.3% of the structured resources rank positively in 3 or more quality attributes, which favors the reuse of them.

4.3.2 Classifying the Structured Resources

For the water quality domain, we calculated the arithmetic average of the normalized values of domain aspects covered in each group by the structured resources and their quality attributes score to compute the final score. The classification was recorded in the “Structured

Resources Classification” table of the dataset [42]. Table 24 shows the ranking for the top 10 structured resources from each group. In some cases, the number of structured resources presented is greater than 10 because more resources were tied in the same position.

Table 24 - Fragment of the Structured Resources Classification

Aspects Group	Structured Resources	Number of Covered Aspects	Number of Covered Aspects Normalized	Quality Attributes Score	Quality Attributes Score Normalized	Final Score
How	INSPIRE	11	1.00	4	0.67	0.83
	O&M	6	0.55	6	1.00	0.77
	GeoSciML	8	0.73	4	0.67	0.70
	ISO/TC 211, ODM	6	0.55	4	0.67	0.61
	SSN	2	0.18	6	1.00	0.59
	USGS Thesaurus	11	1.00	1	0.17	0.58
	GEMET	9	0.82	2	0.33	0.58
Where	Darwin Core, EML	7	0.64	3	0.50	0.57
	GML, ISO/TC 211, WaterML, INSPIRE, UsgsHydroML	3	1.00	4	0.67	0.83
When	Darwin Core, SWEET, GeoNames, TGN, GeoSPARQL, WDTF, GCMD, Upper Cyc	3	1.00	3	0.50	0.75
	O&M	1	1.00	6	1.00	1.00
	OWL-Time	1	1.00	5	0.83	0.92
What	new SSN, SensorML, PROV-O, GML, OM-Lite, SAM-Lite, ISO/TC 211, WaterML, SEGO, INSPIRE, ODM, UsgsHydroML, GeoSciML	1	1.00	4	0.67	0.83
	ISO/TC 211, UsgsHydroML	8	0.89	4	0.67	0.78
	SWEET, EnvO, Upper Cyc	9	1.00	3	0.50	0.75
	QUDT	5	0.56	5	0.83	0.69
	SUMO	9	1.00	2	0.33	0.67
	Uberon, INSPIRE	6	0.67	4	0.67	0.67
	O&M	2	0.22	6	1.00	0.61
	OM	5	0.56	4	0.67	0.61
Who	InAWaterSense, WQOP	8	0.89	2	0.33	0.61
	SSN, O&M	2	0.67	6	1.00	0.83
	SAM-Lite, ISO/TC 211, INSPIRE, UsgsHydroML	3	1.00	4	0.67	0.83
	EML, SWEET	3	1.00	3	0.50	0.75
	new SSN, SensorML, PROV-O, OM-Lite, WaterML, SEGO, ODM, GeoSciML	2	0.67	4	0.67	0.67
Why	MEMOn, ECS	3	1.00	2	0.33	0.67
	INSPIRE, UsgsHydroML	1	1.00	4	0.67	0.83
	SWEET, WDTF, Upper Cyc	1	1.00	3	0.50	0.75
	InAWaterSense, SUMO, PEIA, GEMET	1	1.00	2	0.33	0.67
	USGS Thesaurus, WQO, WaWO+	1	1.00	1	0.17	0.58

As one can observe, some structured resources appear well classified in all or most of the aspects groups. This is the case of INSPIRE, well classified in the 6 groups; ISO/TC 211 and UsgsHydroML, well classified into 5 groups; and O&M and SWEET, well classified into 4 groups.

4.3.3 Evaluating the Structured Resources

We selected 75 elements from five data sources identified in cycle I to be annotated with the structured resources. The data providers are: ANA, IBAMA-IEMA, IGAM, CPRM and

Renova Foundation. The first structured resource evaluated was the INSPIRE since it ranked well in all aspects groups. In its evaluation, 59 of the 75 data sources elements (78.7%) were properly represented. This number indicates that INSPIRE is indeed an artifact to be reused. It is important to inform that 14 (23.7%) of the 59 data sources elements were represented by other structured resources reused by INSPIRE, 12 from O&M and 2 from ISO/TC 2011, also confirming the good positioning of these resources. About the other 16 concepts (21.3%), they are relative to the physical, chemical and biological properties used for water quality measurements. We choose not to represent them with INSPIRE because it treats them very generically. To represent them, we selected QUDT and EnvO, well classified in the What group. QUDT represents each of the properties and units of measure used by the data sources. EnvO represents the chemical entities. It is also important to note that EnvO represents the chemical entities through ChEBI, another resource identified in cycle II, but not ranked so well in the What group because it is focused narrowly on chemical entities. This evaluation is available in the “Structured Resources Evaluation” table of the dataset [42].

Table 25 shows part of this evaluation, focusing on data elements presented in Table 16 of this work. Table 25 contains: the data source, which indicates the provenance of data; the data source element to be annotated; the structured resource that provides the proper representation to the data source element; and the structured resource concept, property and instance that can be used to represent the data source element. For example, in the second row of IGAM, we have the data source element Hydrographic Basin. INSPIRE provides the concept RiverBasin with the property geographicalName to represent it. Another example can be seen in the last row of IBAMA-IEMA that contains the element Alkalinity of bicarbonates (mgCaCO₃/L). The instance Concentration of the concept ChemistryQuantityKind of QUDT is used to represent the chemical property, the concept calcium carbonate of EnvO (ChEBI) is used to represent the chemical entity CaCO₃, the instance MilliGram/Liter of the concept Unit of QUDT is used to represent the unit of measurement, and the concept QuantityValue of QUDT is used to represent the measured value for this chemical property.

In the evaluation performed, we were able to represent all elements of the data sources identified in cycle I with 6 of the structured resources identified in cycle II (INSPIRE, O&M, ISO/TC 2011, QUDT, EnvO and ChEBI). These resources are complementary to each other, with INSPIRE offering broad coverage of domain aspects and the other resources covering some aspects in depth.

Table 25 - Fragment of the Structured Resources Evaluation

Data Source	Data Source		Structured Resource		
	Data Source Element	Name	Concept (class)	Property	Instance
IBAMA- IEMA	Data Provider	INSPIRE	RelatedParty	organisationName	
	Site	INSPIRE	HydroObject / AdministrativeUnit	geographicalName / name	
	Sample Point Short Name	INSPIRE	EnvironmentalMonitoringFacility	name	
	Sample Point Long Name	INSPIRE	EnvironmentalMonitoringFacility	additionalDescription	
	Sample Point Category	INSPIRE	EnvironmentalMonitoringFacility	mediaMonitored	
	Lat	INSPIRE	EnvironmentalMonitoringFacility	representativePoint	
	Long	INSPIRE	EnvironmentalMonitoringFacility	representativePoint	
	X	INSPIRE	EnvironmentalMonitoringFacility	representativePoint	
	Y	INSPIRE	EnvironmentalMonitoringFacility	representativePoint	
	Z	INSPIRE	EnvironmentalMonitoringFacility	representativePoint	
	Projection	INSPIRE (ISO/TC 2011)	CS_CRS		
	Datum	INSPIRE (ISO/TC 2011)	CD_Datum		
	Date	INSPIRE (O&M)	SF_Specimen	samplingTime	
	Sample Ref	INSPIRE (O&M)	SF_Specimen		
	Lab Ref	INSPIRE (O&M)	SF_Specimen		
	Data Source	INSPIRE	RelatedParty	organisationName	
	Sample Type	INSPIRE (O&M)	SF_Specimen	samplingMethod	
	Alkalinity of bicarbonates (mgCaCO3/L)	QUDT	ChemistryQuantityKind		Concentration
		EnvO (ChEBI)	calcium carbonate		
QUDT		Unit		MilliGram/Liter	
QUDT		QuantityValue			
IGAM	Data Provider	INSPIRE	RelatedParty	organisationName	
	Hydrographic Basin	INSPIRE	RiverBasin	geographicalName	
	Sub Basin	INSPIRE	RiverBasin	geographicalName	
	UPGRH	INSPIRE	HydroObject	geographicalName	
	County	INSPIRE	AdministrativeUnit	name	
	Water Course	INSPIRE	Watercourse	geographicalName	
	Description	INSPIRE	EnvironmentalMonitoringFacility	additionalDescription	
	Framing Class of Water Course	INSPIRE	LegislationCitation		
	Station	INSPIRE	EnvironmentalMonitoringFacility	name	
	Altitude	INSPIRE	EnvironmentalMonitoringFacility	representativePoint	
	Latitude (Decimal Degrees)	INSPIRE	EnvironmentalMonitoringFacility	representativePoint	
	Latitude (Degrees Minutes Seconds)	INSPIRE	EnvironmentalMonitoringFacility	representativePoint	
	Longitude (Decimal Degrees)	INSPIRE	EnvironmentalMonitoringFacility	representativePoint	
	Longitude (Degrees Minutes Seconds)	INSPIRE	EnvironmentalMonitoringFacility	representativePoint	
	Year	INSPIRE (O&M)	SF_Specimen	samplingTime	
	Sampling Date	INSPIRE (O&M)	SF_Specimen	samplingTime	
	Sampling Time	INSPIRE (O&M)	SF_Specimen	samplingTime	
Alkalinity of bicarbonates	QUDT	ChemistryQuantityKind			
	QUDT	QuantityValue			

4.4 Related Work

In this section, we investigate reuse-based environmental domain ontologies to verify how the search and selection activities for reusable knowledge resources are addressed in the development of these ontologies. We have analyzed them according with three main aspects: (A1) What was the criterion used to select reusable knowledge resources? (A2) How was domain coverage addressed? (A3) How have the selected knowledge resources been evaluated? This type of related work can be found in the publications selected by the application of CLeAR to the water quality domain, since some of them have also built a shared model (i.e., an ontology) for the environmental domain based on existing structured resources.

For example, in [43] and [44], the authors propose a water quality vocabulary based on knowledge resources such as O&M, QUDT and ChEBI. In [45] and [46], an SSN-based ontology for water quality management (the InAWaterSense ontology) has been developed to support water quality classification based on different regulation authorities. Finally, in [47] the authors propose an ontology-based system with the intent of providing semantic interoperability for environmental monitoring data. As part of this system, an ontology, the Environmental Monitoring Ontology (MEMOn), is developed by reusing others such as SSN, EnvO and the upper level ontology Basic Formal Ontology (BFO) [48].

Regarding aspect A1, in all the works analyzed, the authors did not describe how they found the knowledge resources and objective criteria used to select a specific knowledge resource for reuse. In [43] and [44], the authors justify that QUDT is well-aligned with their understanding of relationships between measurements and units of measure. In [45] and [46], the authors report that SSN ontology is the main upper ontology for modeling WSN (Wireless Sensor Networks) knowledge bases. Thus, this ontology is best suited for the construction of the InAWaterSense core ontology. In [47], the authors report that they have reused some existing ontologies that are relevant for describing environmental monitoring domain such as SSN, EnvO, etc. They explain that they chose these ontologies for two reasons: reduce duplicate work and promote interoperability between ontologies. Besides that, we can see a certain convergence in their choices, probably motivated by the community's acceptance of the knowledge resources.

Concerning aspect A2, none of the works show how the domain is covered by each of the reused knowledge resources. We highlight [47] because it is the only one we identified that adopts a methodology to develop MEMOn and presents how this task was performed. They use an iterative methodology called “Agile methodology for developing Ontology Modules” (AOM). Iterations include defining competency questions, building semi-formal modules, formalizing modules, evaluating modules, and merging modules with other ontological modules. The domain coverage by each module of the final developed ontology is assessed by base metrics (which comprise classes, properties and axioms numbers) and schema metrics (which address the design of the ontology such as inheritance and relationship richness and axiom/class and class/relation ratios).

In relation to the evaluation of the reused knowledge resources (A3), none of the works perform a specific evaluation for them. In [47], a detailed evaluation was performed concerning the final developed ontology which comprises the knowledge resources reused. In this evaluation, the following criteria are considered: (C1) coherence, which refers to the fact that the ontology must not include any contradictions; (C2) interoperability, which represents how the ontology is aligned to upper level or other ontologies; (C3) extensibility, which defines the capability of the ontology to be easily extended by other ontologies; and (C4) completeness, which measures if the domain of interest is appropriately covered by the ontology.

As can be seen, there has been some effort to build ontologies for the environmental domain and reuse has been considered an important factor. Nevertheless, in most related efforts, knowledge resources have been selected with no explicit justification, possibly relying on previous experiences of ontology engineers. Differently, we have proposed that this task be approached systematically, addressing the search process and the criteria to be employed in the selection of knowledge resources for reuse.

4.5 Concluding Remarks

In this chapter, the CLeAR approach has been applied to the water quality domain. We focused on finding structured resources to be reused in the development of the network of reference ontologies for the integration of water quality data. A set of 75 structured resources candidates to be reused were obtained. These knowledge resources were analyzed according to the domain coverage and some quality attributes and classified based on this assessment. In

the evaluation performed, 6 of the structured resources (INSPIRE, O&M, ISO/TC 2011, QUDT, EnvO and ChEBI) were able to jointly represent all elements of the data sources to be integrated. These structured resources were selected to be reused.

It is important to mention that the set of 75 structured resources is available and provides an important knowledge base that can be revisited and reused whenever new needs arise. Thus, people who need to build ontologies for the water quality domain (or environmental domain) can consult it, saving the effort and time required to perform the systematic search and the assessment of the structured resources on this domain.

In an previous work (see [41]), we have conducted a non-systematic search for structured resources about the water quality domain. This non-systematic search resulted in a set of 11 reusable structured resources. Some were already known to us, others were obtained from the analysis of various publications that we can identify. As can be seen, the number of structured resources obtained from the non-systematic search is much lower than the one obtained from the application of the CLeAR approach.

Our impression is that the application of a systematic approach not only guides the work, but also broadens the scope of results and reduces bias. In addition, facilitates discovery of important initiatives and working groups in the field of interest.

In Chapter 5, the six structured resources selected for reuse in the development of the proposed ontology network will be analyzed based on a foundational ontology once they are different and cannot be integrated into their current format.

5 Ontological Analysis of the Knowledge Resources Selected for Reuse

In this chapter, we provide an ontological analysis of the knowledge resources selected for reuse (INSPIRE, O&M, ISO/TC 2011, QUDT, EnvO and ChEBI) based on the Unified Foundational Ontology (UFO) [24][25][26]. The ontological analysis is necessary because these knowledge resources differ from each other in organization, structure, adopted language, etc., and, as a consequence, cannot be integrated into their original format. As the knowledge resources address many elements, we present only those relevant to this work.

By ontological analysis, we mean that knowledge resources elements are classified according to categories of a foundational ontology (UFO in this work). For this, we establish relations between notions of knowledge resources and UFO describing how knowledge resources elements relate to UFO concepts. The relations adopted here were extracted from [49] and are presented by Table 26. As recommended by [49], we have focused on the meanings of each element and concept, instead of on the term used to name them.

Table 26 - Relations used to classify knowledge resources elements according to UFO concepts (extracted from [49])

Relation	Symbol	Meaning
EQUIVALENT	[E]	A is Equivalent to B. Element A represents a notion that is equivalent to the notion represented by Concept B.
SPECIALIZATION OF	[S]	A is a Specialization of B. Element A represents a notion that specializes the notion represented by Concept B.

This chapter is structured as follows. Section 5.1 presents UFO concepts that are required to this work. Section 5.2 focuses on the O&M conceptual model. Section 5.3 discusses the QUDT ontologies. Section 5.4 addresses the INSPIRE conceptual model. Section 5.5 discusses the ISO 19111:2007 Referencing by Coordinates (from ISO/TC 211). Section 5.6 presents and analyses the Environment Ontology (EnvO). Section 5.7 addresses the ChEBI ontology. Finally, section 5.8 presents concluding remarks.

5.1 The Unified Foundational Ontology

The Unified Foundational Ontology (UFO) has been developed based on theories from Formal Ontology, Philosophical Logics, Philosophy of Language, Linguistics and Cognitive Psychology [24][25][26]. UFO consists of three main modules: UFO-A, an ontology of

endurants (objects); UFO-B, an ontology of perdurants (events); and UFO-C, an ontology of social entities built up on UFO-A and UFO-B. The UFO concepts required for this work are presented below.

5.1.1 UFO-A: An Ontology of Endurants

The root concept of UFO is *Entity*, which is specialized into *Universal* and *Individual* [24]. *Individuals* can be *concrete* (e.g., a particular person, an explosion) or *abstract* (e.g., sets, numbers, and propositions). *Concrete Individuals* are divided into *Endurants* and *Perdurants*. *Endurants* are individuals that are wholly present whenever they are present (e.g., a house, a person, an amount of sand, etc.). *Perdurants* are individuals that may have temporal parts. They happen in time in the sense that they extend in time and accumulate temporal parts (e.g., a soccer match). Whenever a perdurant is present, it is not the case that all its temporal parts are present. *Universals* are patterns of features that can be realized in a number of different individuals. Universals can be classified in *Endurant Universals* or *Perdurant Universals*. Endurant universals are endurants patterns of features. Perdurant universals are perdurants patterns of features [24].

UFO-A focuses on endurants (see Figure 9). The category of endurants can be further specialized into *Substantial* and *Moment*. Substantials are existentially-independent individuals (e.g., a house, a person). Moments are individuals that can only exist in other individuals, and, thus, they are existentially-dependent on their bearers (e.g., a color, an electric charge, a social commitment). *Intrinsic Moments* are moments that are dependent on one single individual (e.g., a color, a temperature). *Relators*, in turn, are moments that existentially depend on a plurality of individuals (e.g., an employment, a business process) and, for this reason, provide the material connection between them. *Substantial Universal* and *Moment Universal* are kinds of endurant universals whose individuals are substantials and moments, respectively [24].

Regarding relations, we adopt the *componentOf* relation, which relates individuals that are functional complexes (e.g., a car engine is part of a car, a heart is part of a circulatory system). All parts contribute to the functionality (or the behavior) of the complex [24].

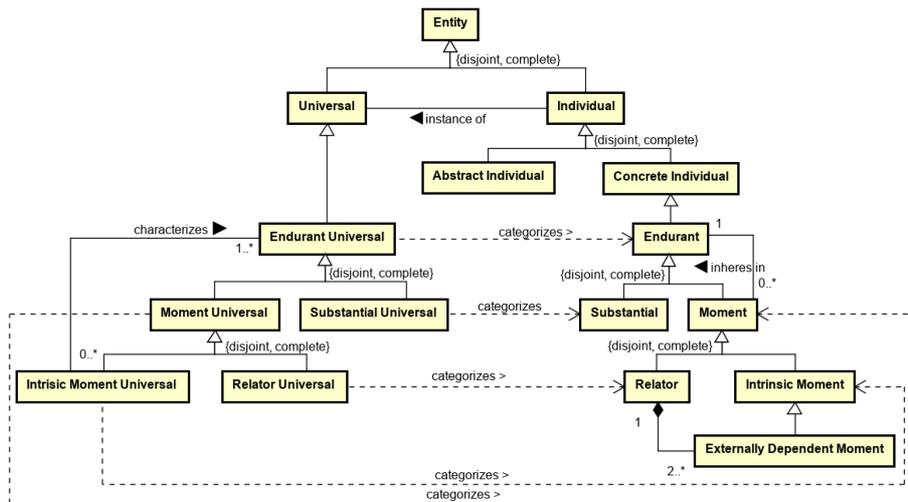


Figure 9 - A fragment of UFO-A [24].

Qualities

Figure 10 presents a fragment of UFO-A related to qualities. Concerning the intrinsic moment universal hierarchy, UFO distinguishes between two main types: *Quality Universals* and *Mode Universals*. Quality universals refer to the properties that characterize universals (e.g., weight, height). They are always associated with values spaces or *Quality Structures* that can be understood as the set of all possible regions (*Quality Regions*) that delimits the space of values that can be associated to a particular quality universal [26]. For example, height and mass are associated with one-dimensional structures with a zero point isomorphic to the half-line of nonnegative numbers. Other properties such as color and taste are represented by multidimensional structures. The perception or conception of an intrinsic moment can be represented as a point in a quality structure. This point is named *Quale*. Quality structures and qualia are together with sets, number and propositions examples of abstract things [25]. *Quality Function*¹ is a specialization of set that maps instances of a quality universal to points in a quality structure [50]. Mode universals are intrinsic moment universals that are not associated with a quality structure (e.g., desire, intention) [24].

¹ In this work, we use the more specific term “Quality Function” to deal with what is called “Function” in [50].

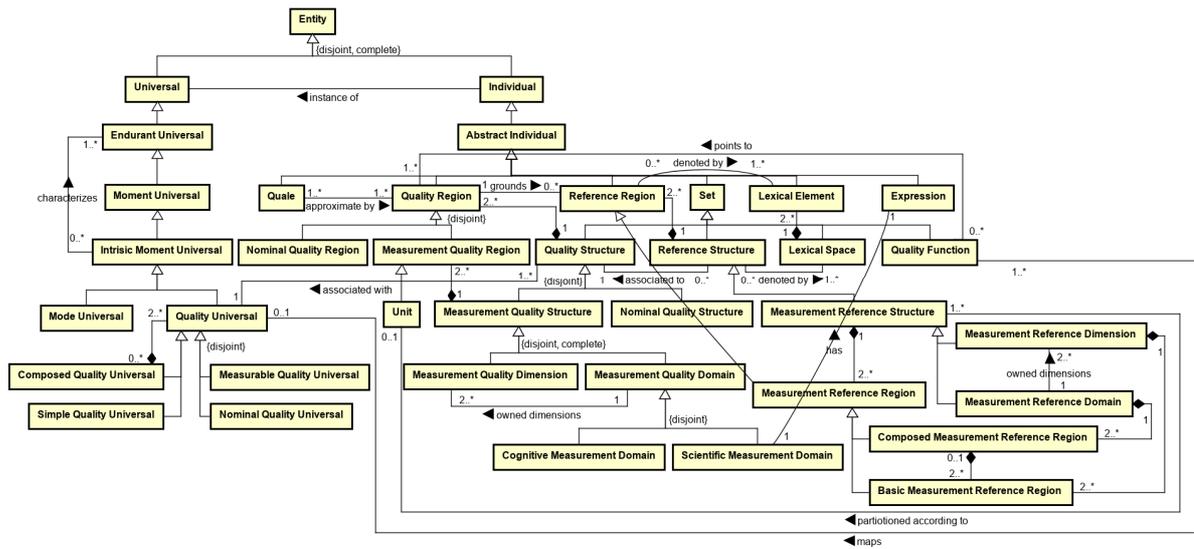


Figure 10 - A fragment of UFO-A related to Qualities [24][25][26][50].

According to the quality structures to which they are associated, quality universals are classified into *Simple* and *Composed Quality Universals*. The first one is associated to one-dimensional quality structures and the last one is associated to multidimensional structures. Regarding their nature, quality universals are classified into *Measurable Quality Universals* and *Nominal Quality Universals*. Measurable quality universals are quality universals that can be objectively measured by cognitive agents or measurement devices, and it is possible to establish distances among their quality regions (e.g., length, height, temperature). Differently, nominal quality universals are usually based on social conventions and cannot be objectively measured (e.g., name, zip code) [26].

Quality structures are divided into *Measurement Quality Structures* and *Nominal Quality Structures*. Measurement quality structures are structures that allow for objectively evaluating the distance between two values and verifying if the values are equal or not. They are classified, according to the number of dimensions, into *Measurement Quality Dimension* and *Measurement Quality Domain*. The first one represents the most elementary (one-dimensional) measurement quality structures, and the last one represents multidimensional quality structures [26].

Measurement quality domains can be *Cognitive Measurement Quality Domain* or *Scientific Measurement Quality Domain*. The practical difference between them is that regions from scientific domains can be qualitatively evaluated and ordered, while regions from cognitive domains cannot. Scientific domains are composed following some kind of

algebra and have an *Expression* that determines their formation. For example, the scientific domain for the body mass indicator (BDI) is formed using the dimensions weight and height ($BMI = \text{weight} / (\text{height} \times \text{height})$) [26].

A quale is intrinsic to cognitive agents and therefore cannot be shared or communicated. In order to allow quale communication, it is necessary to use *Lexical Elements* (e.g., 1.86 can be the lexical element used to communicate the height of a person) associated to *Reference Regions* and *Reference Structures*. A reference region is an abstract thing based on a quality region that acts as a bridge between that region and the lexical elements used to communicate the approximated quale. A reference structure, in turn, is associated to a quality structure and is a set of reference regions grounded in quality regions of that quality structure. When the ‘value’ of a particular quality is being referred by lexical elements (e.g., 1.86), what is actually being referred is a quality region that most approximates the quale [26].

Reference structures are topologically isomorphic to the quality structures to which they are associated. Thus, they have the same number of dimensions and their reference regions are isomorphic to the quality regions of the quality structure. Reference structures associated to measurement quality structures are called *Measurement Reference Structures* (specialized into *Measurement Reference Dimension* and *Measurement Reference Domain*) and act like scales grounded by quality structures. They are composed by *Measurement Reference Regions* (specialized into *Basic Measurement Reference Region* and *Composed Measurement Reference Region*). Measurement reference structures can be partitioned in spaces with the same magnitude according to a *Unit* [50].

5.1.2 UFO B: An Ontology of Perdurants

As presented in Figure 11, UFO-B focuses on perdurants. The main category of UFO-B is *Event*. Events can be atomic or complex. *Atomic Events* have no proper parts. *Complex Events* are aggregations of at least two disjoint events, which can also be atomic or complex. Events are ontologically dependent entities in the sense that they depend on substantial participation to exist. Take for instance the event of measuring the height of a person. In this event, we have the participation of the measured person, the person that performs the measurement and the instrument used to measure the height. This event is composed of the individual participation of each of these entities and depends on them to exist. Besides that, each event is

associated with two *Time Points*: a begin and an end time point. Time points are abstract individuals strictly ordered by a precedes relation [24][25].

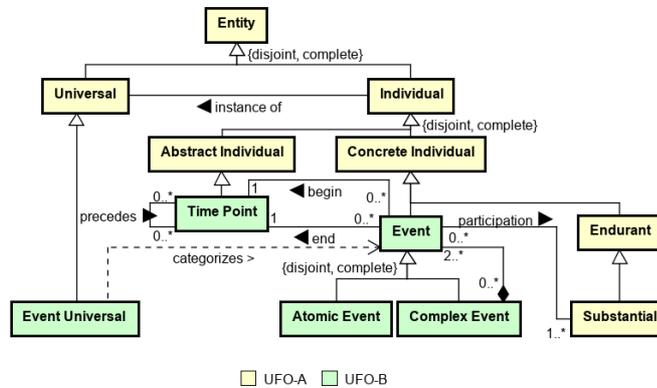


Figure 11 - A fragment of UFO-B [24][25].

5.1.3 UFO C: An Ontology of Social Entities

UFO-C is an ontology of social entities (both endurants and perdurants). A fragment of this ontology is shown in Figure 12. A basic distinction in UFO-C is between agentive and non-agentive substantial individuals, termed *Agents* and *Objects*, respectively. Agents can be divided into *Physical Agents* (e.g., a person) and *Social Agents* (e.g., an organization, a society). Objects can also be further categorized in *Physical Objects* and *Social Objects*. Physical objects include a book, a car, among others; social objects include money, language, etc. A *Normative Description* is a type of social object that defines one or more rules/norms recognized by at least one social agent. Examples of normative descriptions include contracts in general, but also sets of directives on how to perform actions within an organization [25].

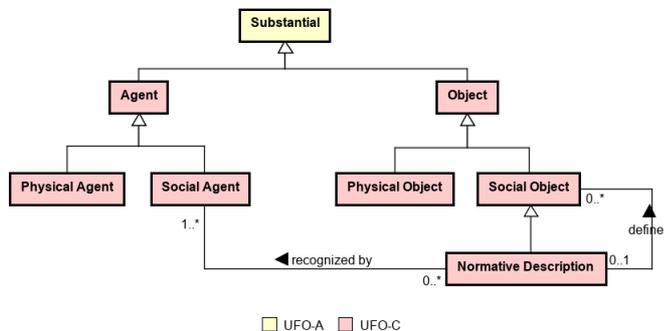


Figure 12 - A fragment of UFO-C related to agents, objects and normative descriptions [25].

5.2 The O&M Conceptual Model

The ISO 19156:2011 Observations and Measurements Standard (O&M) [51] defines a conceptual schema, using the Unified Modeling Language (UML), for observations and for features involved in samplings related to observations. According to O&M, an observation is an act associated with a discrete time instant or period through which a number, term or other symbol is assigned to a phenomenon (property of a feature). It involves the application of a specified procedure, such as a sensor, instrument, algorithm or process chain. The procedure may be applied in-situ, remotely, or ex-situ with respect to the sampling location. Generally, the term “measurement” is used. However, a distinction between measurement and category-observation has been adopted so the term “observation” is used for the general concept. “Measurement” may be reserved for cases where the result is a numeric quantity.

5.2.1 Overview of the O&M Conceptual Model

Figure 13 presents the UML class diagram for the basic *Observation* type (OM_Observation) extracted from [51]. An observation is conceived as an event that aims to measure or otherwise determine the value of a *property* (GFI_PropertyType) of a *feature of interest* (GFI_Feature), using a specified *procedure* (OM_Process) [51].

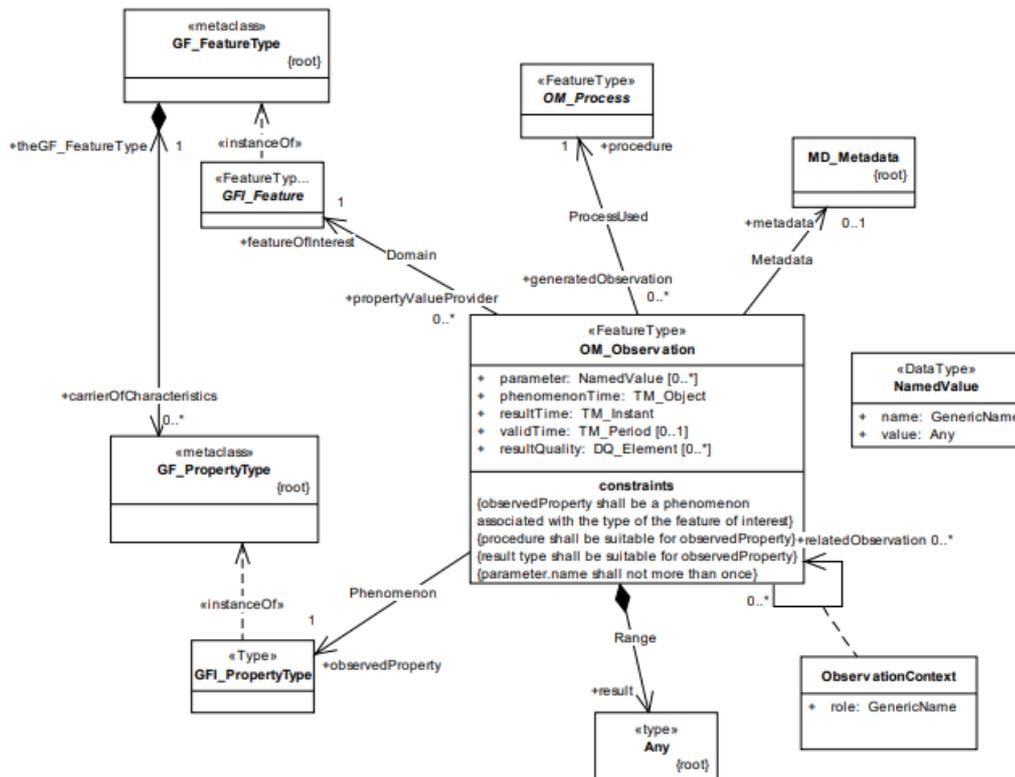


Figure 13 - The basic Observation type extracted from [51].

Multiple temporal attributes are provided for the observation type. The *phenomenonTime* describes the time that the result applies to the property of the feature of interest. This is often the time of interaction by a sampling procedure or observation procedure with a real-world feature. The *resultTime* deals with the time when the result became available, typically when the procedure associated with the observation was completed. For some observations this is identical to the phenomenon time. However, there are important cases where they differ. For example, when a measurement is made on a specimen (physical sample) in a laboratory, the phenomenon time is the time the specimen was retrieved from its host, while the result time is the time the laboratory procedure was applied. The *validTime* describes the time period during which the result is intended to be used. This attribute is commonly required in forecasting applications [51].

The attribute *parameter* deals with an arbitrary event-specific parameter. This might be an environmental parameter, an instrument setting or input, or an event-specific sampling parameter that is not tightly bound to either the feature or to the observation procedure. The *resultQuality*, an instance-specific description, complements the description of the observation procedure, which provides information concerning the quality of all observations using this procedure [51].

Specializations of the observation class have been classified by the result type. For example a *Measurement* is an observation whose result is a scaled quantity (or measure), and a *TruthObservation* is an observation whose result is a Boolean value [51].

Most observations are actually made on representative samples of the feature of interest, so a model of features used for sampling was developed as separate part of O&M. A sampling feature is a feature constructed to support the observation process, which may or may not have a persistent physical expression but would either not exist or be of little interest in the absence of an intention to make observations [51]. Figure 14 presents the UML class diagram for the *SamplingFeature* core (SF_SamplingFeature) extracted from [51].

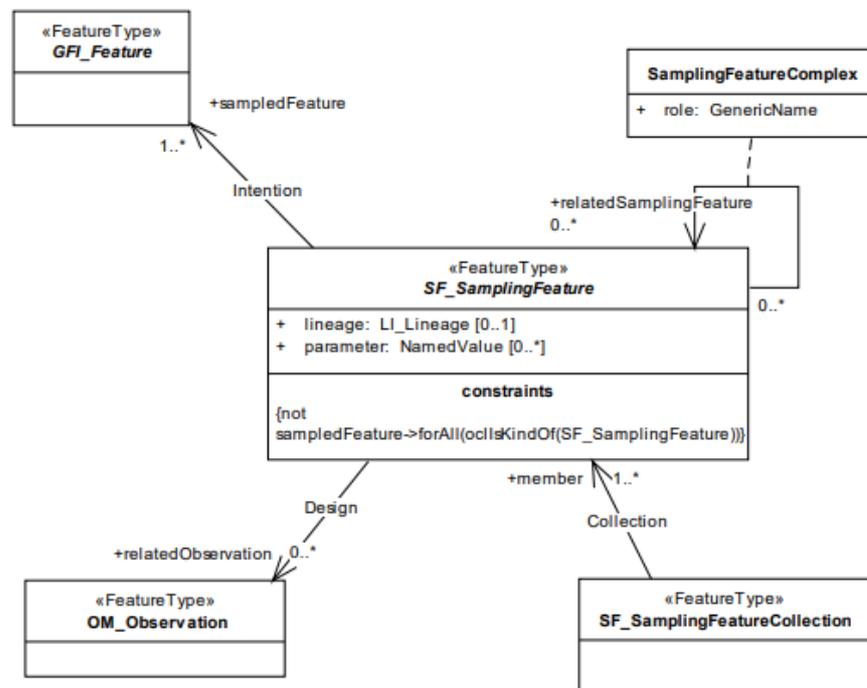


Figure 14 - The SamplingFeature core extracted from [51].

The essential property of a generic sampling feature is the *sampledFeature* relationship with the feature that it samples. A profile typically samples a water or atmospheric column; a well samples the water in an aquifer; a tissue specimen samples a part of an organism. The attribute *parameter* of this class describes an arbitrary parameter associated with the sampling feature. This might be a parameter that qualifies the interaction with the sampled feature (GFI_Feature), or an environmental parameter associated with the sampling process. The *lineage* deals with the history and provenance of the sampling feature. This might include information relating to the handling of the specimen, or details of the survey procedure of a spatial sampling feature [51].

A specimen is a physical sample, obtained for observation(s) carried out ex situ, sometimes in a laboratory [51]. The *Specimen* (SF_Specimen) UML class diagram extracted from [51] is shown in Figure 15.

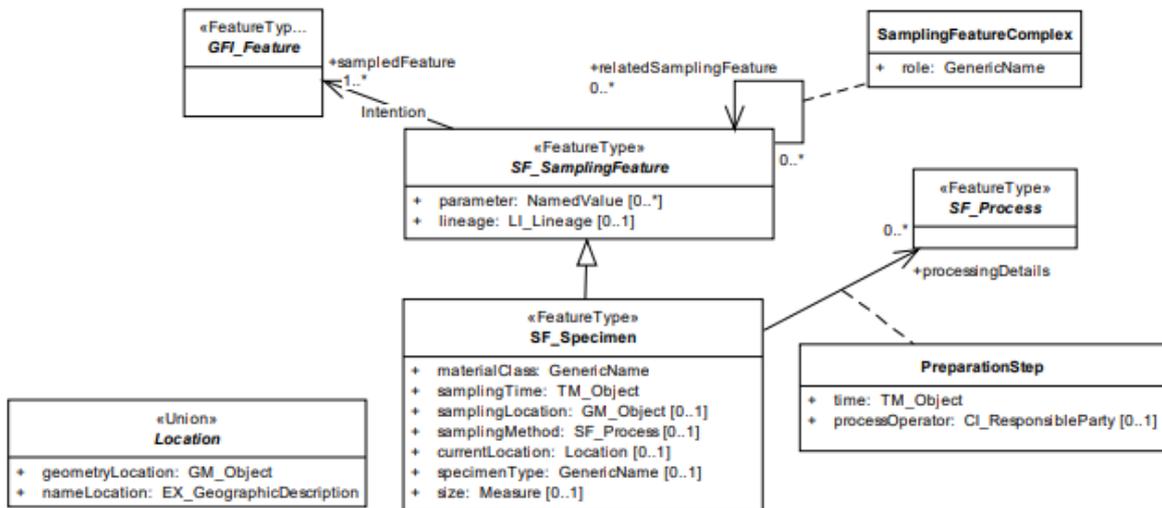


Figure 15 - The Specimen model extracted from [51].

With regard to a specimen, the attribute *materialClass* provides a basic classification of the material type of the specimen (e.g., soil, water, rock, vegetation). The *samplingMethod* describes the method used to obtain the specimen from its sampled feature. The *samplingTime* records when the specimen was retrieved from the sampled feature. The *samplingLocation* describes the location from where the specimen was obtained. The attribute *currentLocation* deals with the location of a physical specimen. This may be a storage location, such as a shelf in a warehouse or a drawer in a museum. The *specimenType* describes the basic form of the specimen (e.g., polished section, core, pulp, solution). The attribute *size* describes a physical extent of the specimen. This may be length, mass, volume, etc. as appropriate for the specimen instance and its material class [51].

In many applications, specimen preparation procedures are applied to the material prior to its use in an observation. The class *PreparationStep* links a specimen to a process that describes a phase of the specimen preparation. The attribute *time* of this class describes the time that the process was applied to the specimen. It supports ordering of preparation steps. The *processOperator* is related to the operator (responsible party) of the process involved in the preparation step [51].

5.2.2 Ontological Analysis of the O&M Conceptual Model

In O&M, *Observation*, *Measurement*, *TruthObservation*, other observation specializations, and *PreparationStep* are events. As a consequence, they are classified as specializations of

UFO-B Event. *Feature* is an abstraction of real-world phenomena (including both objects and events) and is classified as a specialization of UFO-A Individual. *PropertyType* is a type of characteristic of a feature and is equivalent to UFO-A Quality Universal.

An observation *procedure* can be a method, algorithm or instrument. We can see that this element mixes concepts from different UFO categories. When dealing with methods and algorithms, it is classified as a specialization of UFO-C Normative Description. When dealing with instruments, it is classified as a specialization of UFO-C Physical Object. In turn, an observation *result* is classified as a specialization of UFO-A Abstract Individual once it represents the value of any property.

Regarding observation temporal attributes, they are classified as specializations of UFO-A Abstract Individual. Particularly, *resultTime* is related to observation begin and end time points. Thus, result time refers to specializations of UFO-B Event begin and end Time Points. In turn, *phenomenonTime* is related to sampling begin and end time points and *validTime* is related to the period to which simulations apply, but sampling and simulation are not explicitly modeled.

The attribute *parameter* of an observation can be many different things (an environmental parameter, an instrument setting or input, etc.). Therefore, we will not classify it into a UFO category. In the ontology network proposed in the next chapter, each relevant property of a research activity (procedure and instrument adopted, agents involved, etc.) must be explicitly modeled. As a consequence, this attribute will not be reused. In turn, the attribute *resultQuality* is a description of the observation result and refers to a specialization of UFO-A Abstract Individual.

SamplingFeature is a feature, such as a station, transect, section or specimen, which is involved in making observations concerning a domain feature. Then, it can be classified as a specialization of UFO-A Individual. A *sampledFeature* is a feature too and is also classified as a specialization of UFO-A Individual. As for the attribute *parameter* of an observation, the *parameter* of a sampling feature can be many different things and will not be classified into a UFO category (following our strategy for *parameter* of an observation as discussed above). Since *lineage* is an unstructured attribute (a string), it will not be reused.

Specimen is a physical sample. As a consequence, it is classified as a specialization of UFO-A Substantial. The attributes *specimenType* and *materialClass* classify the specimen and the specimen type, respectively. Thus, they refer to specializations of UFO-A Substantial Universal. As the *samplingMethod* can be a method or instrument, it refers to a specialization of UFO-A Substantial. In turn, the attribute *samplingTime* refers to a specialization of UFO-A Abstract Individual. It is related to sampling begin and end time points, but sampling was not explicitly modeled. The attributes *samplingLocation* and *currentLocation* describe spatial location. They refer to specializations of UFO-A Abstract Individual. The attribute *size* is a measure and refers to a specialization of UFO-A Abstract Individual too.

With regards preparation, the attribute *time* refers to specializations of UFO-B Event begin and end Time Points. The attribute *processOperator* refers to a specialization of UFO-C Agent. Table 27 summarizes the relations between O&M Conceptual Model elements and UFO concepts.

Table 27 - Relations between O&M Conceptual Model elements and UFO concepts

O&M Conceptual Model element	Relation Symbol	UFO concept
Observation	[S]	UFO-B: Event
Measurement, TruthObservation, other observation specializations	[S]	UFO-B: Event
PreparationStep	[S]	UFO-B: Event
Feature (feature of interest, sampled feature)	[S]	UFO-A: Individual
PropertyType	[E]	UFO-A: Quality Universal
procedure (method, algorithm)	[S]	UFO-C: Normative Description
procedure (instrument)		UFO-C: Physical Object
result	[S]	UFO-A: Abstract Individual
resultTime	[S]	UFO-B: begin and end Time Points
phenomenonTime	[S]	UFO-A: Abstract Individual
validTime	[S]	UFO-A: Abstract Individual
resultQuality	[S]	UFO-A: Abstract Individual
SamplingFeature	[S]	UFO-A: Individual
Specimen	[S]	UFO-A: Substantial
specimenType	[S]	UFO-A: Substantial Universal
materialClass	[S]	UFO-A: Substantial Universal
samplingMethod	[S]	UFO-A: Substantial
samplingTime	[S]	UFO-A: Abstract Individual
samplingLocation	[S]	UFO-A: Abstract Individual

currentLocation	[S]	UFO-A: Abstract Individual
size	[S]	UFO-A: Abstract Individual
time	[S]	UFO-B: begin and end Time Points
processOperator	[S]	UFO-C: Agent

5.3 The QUDT Ontologies

The “Quantity, Unit, Dimension and Type” (QUDT) ontologies [52] define the base classes, properties and restrictions used for modeling physical quantities, units of measure, and their dimensions in various measurement systems. The goal of the QUDT collection of models is to provide a unified model of measurable quantities, units for measuring different kinds of quantities, the numerical values of quantities in different units of measure and the data structures and data types used to store and manipulate these objects in software. QUDT ontologies are organized as collections of different types of graphs, as listed in the QUDT catalog. Vocabulary graphs hold different domains of quantities and units, which import the appropriate QUDT schemas (RDF/OWL).

5.3.1 Overview of the QUDT Ontologies

Figure 16 shows the conceptual model of QUDT ontologies extracted from [52]. Below, the main classes of this model are explained.

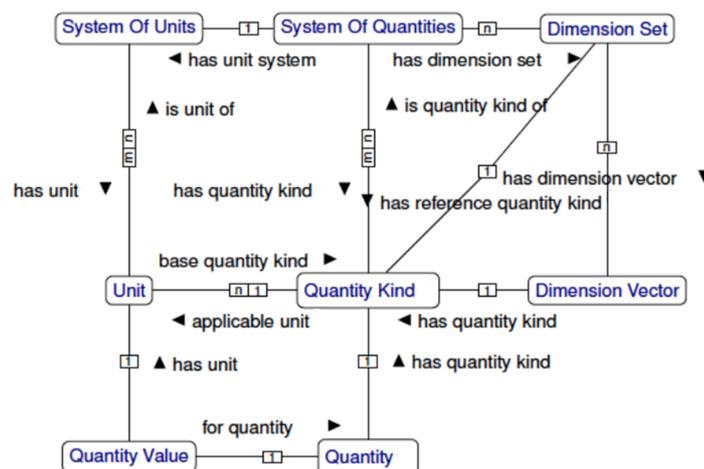


Figure 16 - Conceptual Model of QUDT extracted from [52].

A *Quantity Kind* is any observable property that can be measured and quantified numerically. Familiar examples include physical properties such as length, mass, time, force, energy, power, electric charge, etc. Less familiar examples include currency, interest rate,

price to earnings ratio, and information capacity. *Derived Quantity Kinds* are defined in terms of a small set known as *Base Quantity Kinds* using physical laws [52].

A *Quantity* is defined in [52] as the “measurement of an observable property of a particular object, event, or physical system”. Quantities are differentiated by two attributes which together comprise the essential parameters needed to characterize what is measured: kind and magnitude. The kind of a quantity identifies the observable property quantified (e.g., length, force, frequency); the magnitude of a quantity expresses its relative size compared to other quantities of the same kind. For example, the speed of light in a vacuum and the escape velocity of the Earth are both quantities of the kind speed but are of different magnitudes [52].

A unit of measurement, or *Unit*, is a particular quantity of a given kind that has been chosen as “a scale for measuring other quantities of the same kind” [52]. For example, the Meter is a quantity of length that has been empirically defined and standardized by the International Board of Weights and Measures (BIPM). Any quantity of length can be expressed as a number multiplied by the unit meter. More formally, the value of a quantity Q with respect to a unit U is expressed as the scalar multiple of a real number n and U , as $Q = nU$ [52].

A *Quantity Value* expresses the numerical value of a quantity with respect to a chosen unit of measurement. For example, the value of Planck’s constant in Joule-Seconds (J s) is approximately 6.62606896E-34, whereas the value in Erg-Seconds (erg s) is approximately 6.62606896E-27 [52].

A *System of Quantities* is a specification, typically developed and maintained by an authoritative source, of the base quantity kinds for the system; and the formulas expressing each derived quantity kind in the system in terms of the base quantity kinds. For example, the International System of Quantities (ISQ) is used with the International System of Units (SI) [52].

A *System of Units* is a choice of base units and derived units, together with their multiples and submultiples, defined in accordance with given rules, for a given system of quantities. A *Base Unit* is a unit of measurement for a base quantity. A *Derived Unit* is a unit of measurement for a derived quantity [52].

A *Dimension Vector* is an expression of the dependence of a quantity on the base quantity kinds of a system of quantities as a product of powers of factors corresponding to the base quantities, omitting any numerical factor. For instance, the dimension of the physical quantity *speed* is *length/time*, and the dimension of the physical quantity *force* is *mass x acceleration* or *mass x (length/time)/time* [52].

5.3.2 Ontological Analysis of the QUDT Ontologies

In QUDT Ontologies, *Quantity Kinds* are any observable property that can be measured and quantified numerically. They are equivalent to UFO-A Measurable Quality Universals. *Base* and *Derived Quantity Kinds* are equivalent to simple and composed measurable quality universals, respectively.

Quantities are characterized by two attributes: kind and magnitude. As the kind attribute identifies the observable property quantified, it refers to a UFO-A Quality Universal. In turn, the magnitude attribute (i.e., the *Quantity Value*) represents a value associated to a quality in a particular context of measurement, and thus it refers to a UFO-A Measurement Reference Region. A *Unit* is a particular quantity of a given kind that has been chosen as “a scale for measuring other quantities of the same kind”. As a consequence, units have the same attributes as quantities, and these attributes have the same classification with respect to UFO categories as quantities attributes. More specifically, a *Unit’s* kind refers to a UFO-A Quality Universal, and *Unit’s* magnitude refers to a UFO-A Unit (a specialization of UFO-A Measurement Quality Region).

Systems of Quantities and *Systems of Units* are classified as specializations of UFO-C Normative Description. *Dimension Vectors* are classified as specializations of UFO-A Expression. Table 28 presents the relations between the QUDT Ontologies elements and UFO concepts.

Table 28 - Relations between the QUDT Ontologies elements and UFO concepts

QUDT Ontologies element	Relation Symbol	UFO concept
Quantity Kind	[E]	UFO-A: Measurable Quality Universal
Base Quantity Kind	[E]	UFO-A: Simple Measurable Quality Universal
Derived Quantity Kind	[E]	UFO-A: Composed Measurable Quality Universal
Quantity (kind)	[E]	UFO-A: Measurable Quality Universal
Quantity (magnitude)	[E]	UFO-A: Measurement Reference Region
Unit (kind)	[E]	UFO-A: Measurable Quality Universal
Unit (magnitude)	[E]	UFO-A: Unit (a specialized Measurement Quality Region)
Quantity Value	[E]	UFO-A: Measurement Reference Region
System of Quantity	[S]	UFO-C: Normative Description
System of Unit	[S]	UFO-C: Normative Description
Dimension Vector	[S]	UFO-A: Expression

5.4 The INSPIRE Conceptual Model

INSPIRE (Infrastructure for Spatial Information in Europe) [53] is a European Union spatial data infrastructure for the purposes of EU environmental policies and policies or activities which may have an impact on the environment. This European Spatial Data Infrastructure aims to enable the sharing of environmental spatial information among public sector organizations, facilitate public access to spatial information across Europe and assist in policy-making across boundaries. INSPIRE is based on the infrastructures for spatial information established and operated by the Member States of the European Union. It addresses 34 spatial data themes needed for environmental applications, such as hydrography, transport networks, land cover, land use, atmospheric conditions, and environmental monitoring facilities, among others.

5.4.1 Overview of the INSPIRE Conceptual Model

In this work, we are interested in the following themes addressed by INSPIRE: hydrography, administrative units, and environmental monitoring facilities. Next, we present an overview of the relevant concepts of the UML model of each one of them.

Overview of the Hydrography UML Model

In Hydrography, the “Hydro - Physical Waters” conceptual schema defines spatial object types for a range of real-world physical feature classes having a strong relationship to

hydrography [54]. Figure 17 shows the UML class diagram for this conceptual schema extracted from [54].

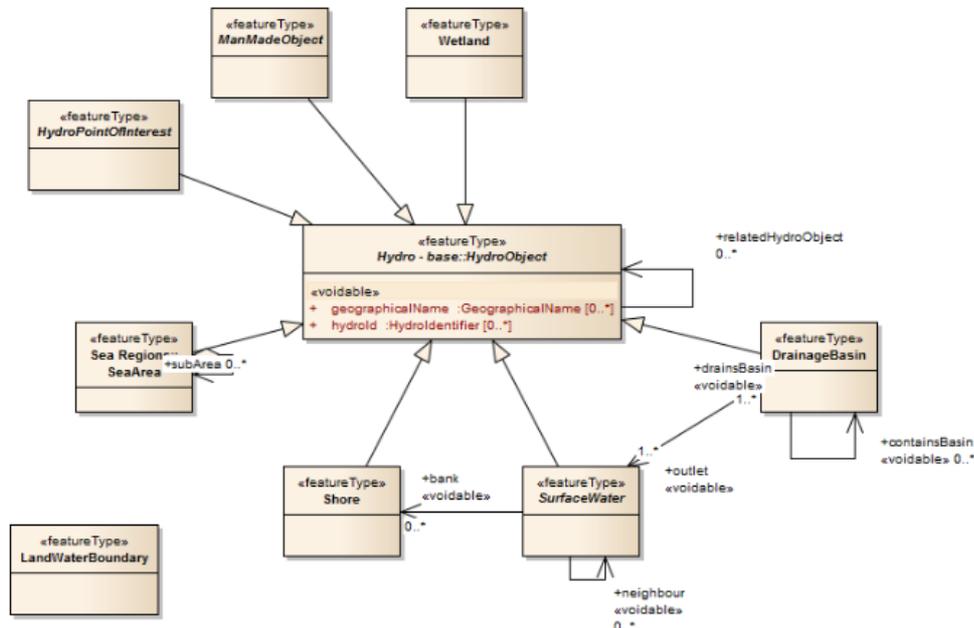


Figure 17 - UML Class Diagram for the “Hydro - Physical Waters” conceptual schema extracted from [54].

The abstract spatial object *HydroObject* is used as a base for hydrographic (including man-made) objects in the real world. The *geographicalName* attribute is a name used to identify a hydrographic object in the real world. The *hydroId* attribute is an identifier that is used to identify a hydrographic object in the real world. More than one identifier may be required, for instance a watercourse may be assigned to different identifying codes under national and European schemes [54].

DrainageBasin represents an area having a common outlet for its surface runoff. Regarding the different classifications of drainage basins, no distinction is made between drainage basins/sub-basins since this will vary with application. It is possible to build basins from other basins. The outlet of a drainage basin may be a canal or a lake. Synonyms for drainage basin include: catchment; catchment area; drainage area; river basin; watershed [54].

The abstract object *SurfaceWater* deals with any known inland waterway body such as lake/pond, reservoir, river/stream, etc. Surface water is related to one or more drainage basins drained by it. *SurfaceWater* can be specialized in *Watercourse* that is a natural or man-made flowing watercourse or stream [54].

ManMadeObject represents an artificial object which lies inside a body of water and has one of the following types of function: retains the water; regulates the quantity of water; alters the course of the water; allows watercourses to cross each other. Examples of this object are embankment, dam or weir, crossing, among others [54].

Lastly, *SeaArea* is an area of sea defined according to its physical and chemical characteristics. It includes named seas such as “Baltic Sea” and also un-named areas of sea that have particular chemical and physical characteristics [54].

Overview of the Administrative Unit UML Model

Figure 18 shows the *AdministrativeUnit* spatial object extracted from [55]. It represents administrative units at all levels of administrative hierarchy. Each single unit (i.e., instance of *AdministrativeUnit* spatial object type) is associated to exactly one hierarchy level. Information about the level in the respective national hierarchy is documented by the mandatory attribute *nationalLevel* [55].

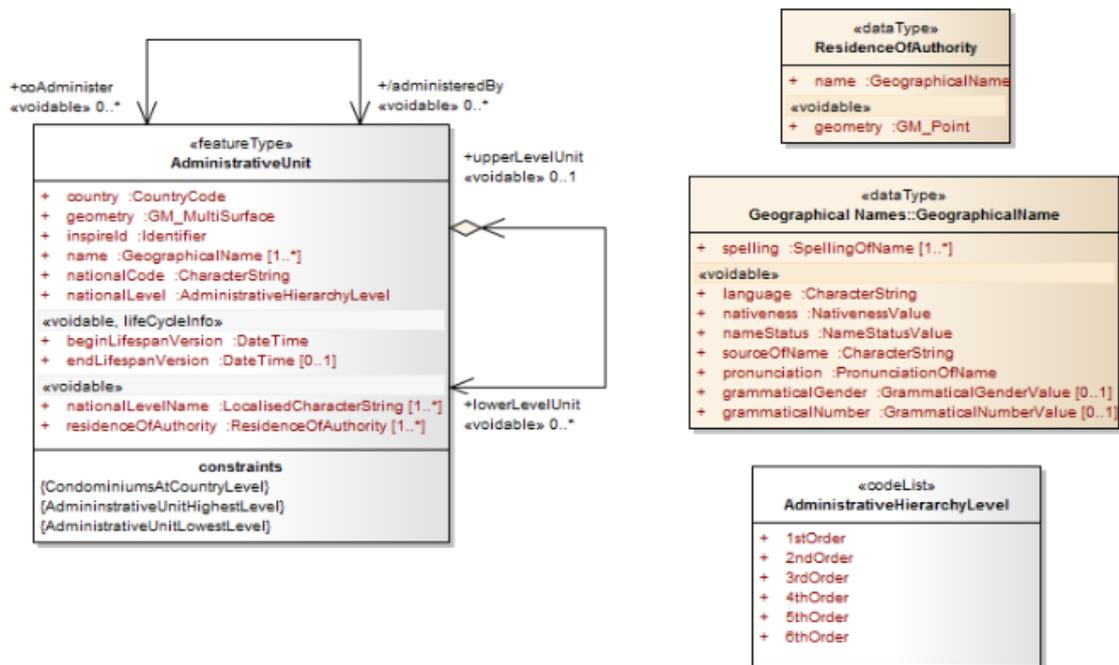


Figure 18 - UML Class Diagram for the “AdministrativeUnit” spatial object extracted from [55].

The number of administrative levels differs from country to country; therefore no absolute levels can be fixed. Instead, the (spatial) correspondence between the levels is a common characteristic of national administrative hierarchies. The representation of these relationships between the units is supported in this conceptual schema by a self-reference of

the *AdministrativeUnit* type, and corresponding to the *lowerLevelUnit* and *upperLevelUnit* association roles [55].

In some countries the hierarchy of administrative units differs from the ideal strictly hierarchical organization. For instance, some units (at lowest level) are not linked to any unit at a higher level but to two or more units at same level. In order to support such situations a self-reference of *AdministrativeUnit* with the *coAdminister* and *administeredBy* association roles is established in this conceptual schema [55].

The attribute *country* is a two-character country code according to the Interinstitutional style guide published by the Publications Office of the European Union. It is used to identify the country to which an administrative unit belongs. The attribute *geometry* is a geometric representation of the spatial area covered by the administrative unit. The attribute *name* is an official national geographical name of the administrative unit. The *nationalCode* is a thematic identifier corresponding to the national administrative codes defined in each country. The *nationalLevelName* is a name of the level in the national administrative hierarchy, at which the administrative unit is established [55].

Overview of the Environmental Monitoring Facilities UML Model

The “Environmental Monitoring Facilities” application schema includes two aspects. The environmental monitoring facility as a spatial object, and observations and measurements linked to the environmental monitoring facility [56]. Figure 19 presents the UML class diagram for the first aspect extracted from [56].

The *EnvironmentalMonitoringFacility* (*EMF*) is the central spatial object type for both aspects. An EMF is a georeferenced object directly collecting or processing data about objects whose properties (e.g., physical, chemical, biological or other aspects of environmental conditions) are repeatedly observed or measured. An EMF can also host other environmental monitoring facilities. Thus, the model provides a recursive hierarchical link (*relatedTo*) between EMFs. This reflects the fact that a station can have various parts or a platform can host a number of sensors or measuring equipment [56].

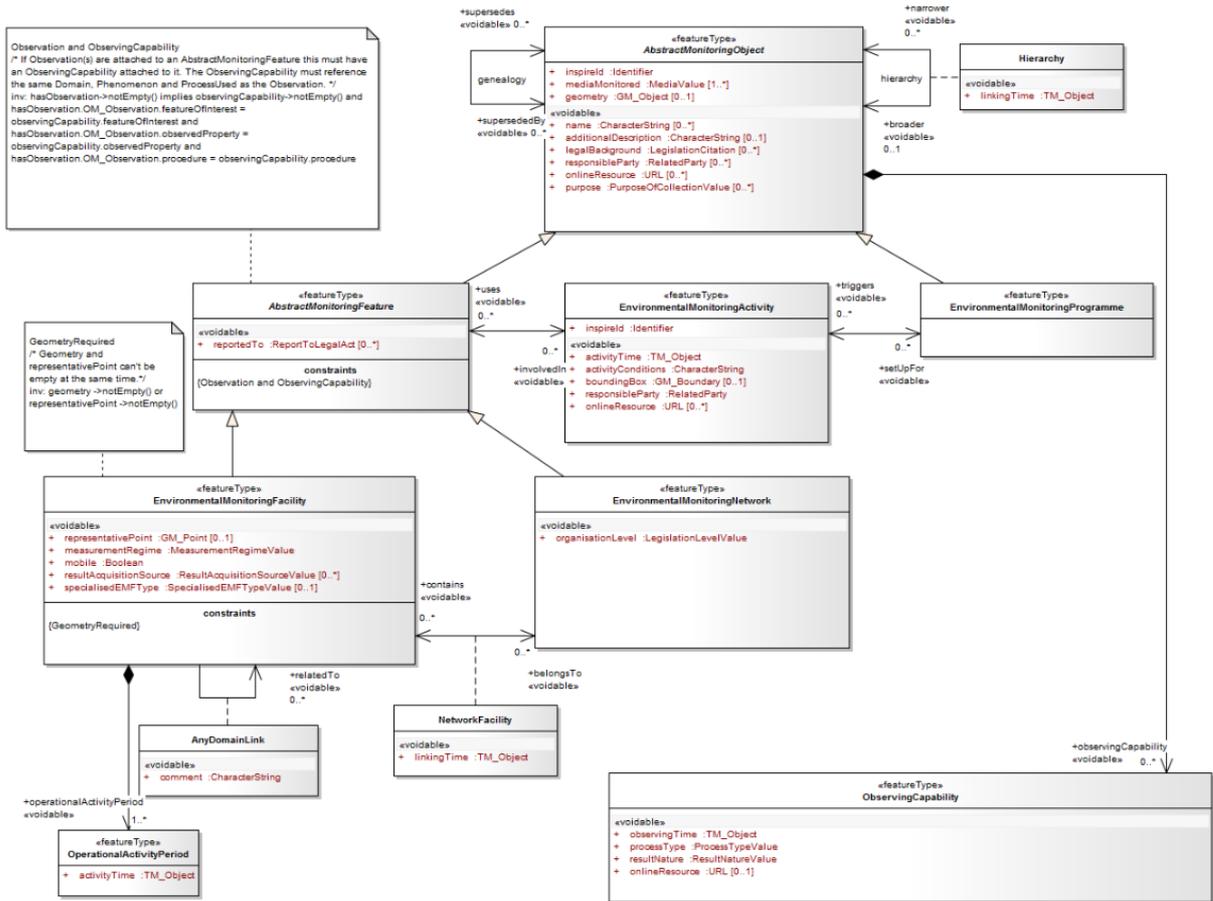


Figure 19 - Fragment of the UML Class Diagram for the “Environmental Monitoring Facilities” conceptual schema related to environmental monitoring facilities extracted from [56].

An EMF includes the attribute *representativePoint* to have a representative location in thematic contexts. The attribute *measurementRegime* of EMF represents the regime of measurement. The *specialisedEMFType* categorizes EMFs as platform, site, station, sensor, etc. The *operationalActivityPeriod* is related to the lifespan of the physical object (facility). With regards to the attributes inherited from *AbstractMonitoringObject*, *mediaMonitored* represents the monitored environmental medium, such as water, air, etc. The *responsibleParty* is the responsible party for the facility [56].

An *EnvironmentalMonitoringProgramme (EMP)* is a policy relevant description defining the target of a collection of observations and/or the deployment of environmental monitoring facilities in the field. Usually an EMP has a long-term perspective over at least a few years. An EMP covers an area of interest (e.g., a region) and is based on environmental legislation. The attributes *geometry* and *legalBackground*, inherited from *AbstractMonitoringObject*, represent the geometric area associated to the facility and the legislation that regulates the facility, respectively [56].

The *EnvironmentalMonitoringActivity (EMA)* expresses the need to describe environmental monitoring campaigns which are carried out with specific equipment for a specific period of time. Examples could be a research cruise of a vessel with monitoring equipment in the ocean or flights by an airplane hosting various sensors for airborne observations. These examples show the high relevance for mobile environmental monitoring facilities in relation to a long term perspective of environmental monitoring programmes. The attribute *activityTime* represents the lifespan of the EMA. The *responsibleParty* represents the responsible party for the EMA. The association role *uses* specifies the set of EMFs involved in an EMA. The association role *setUpFor* specifies the EMPs for which the EMA is set up [56].

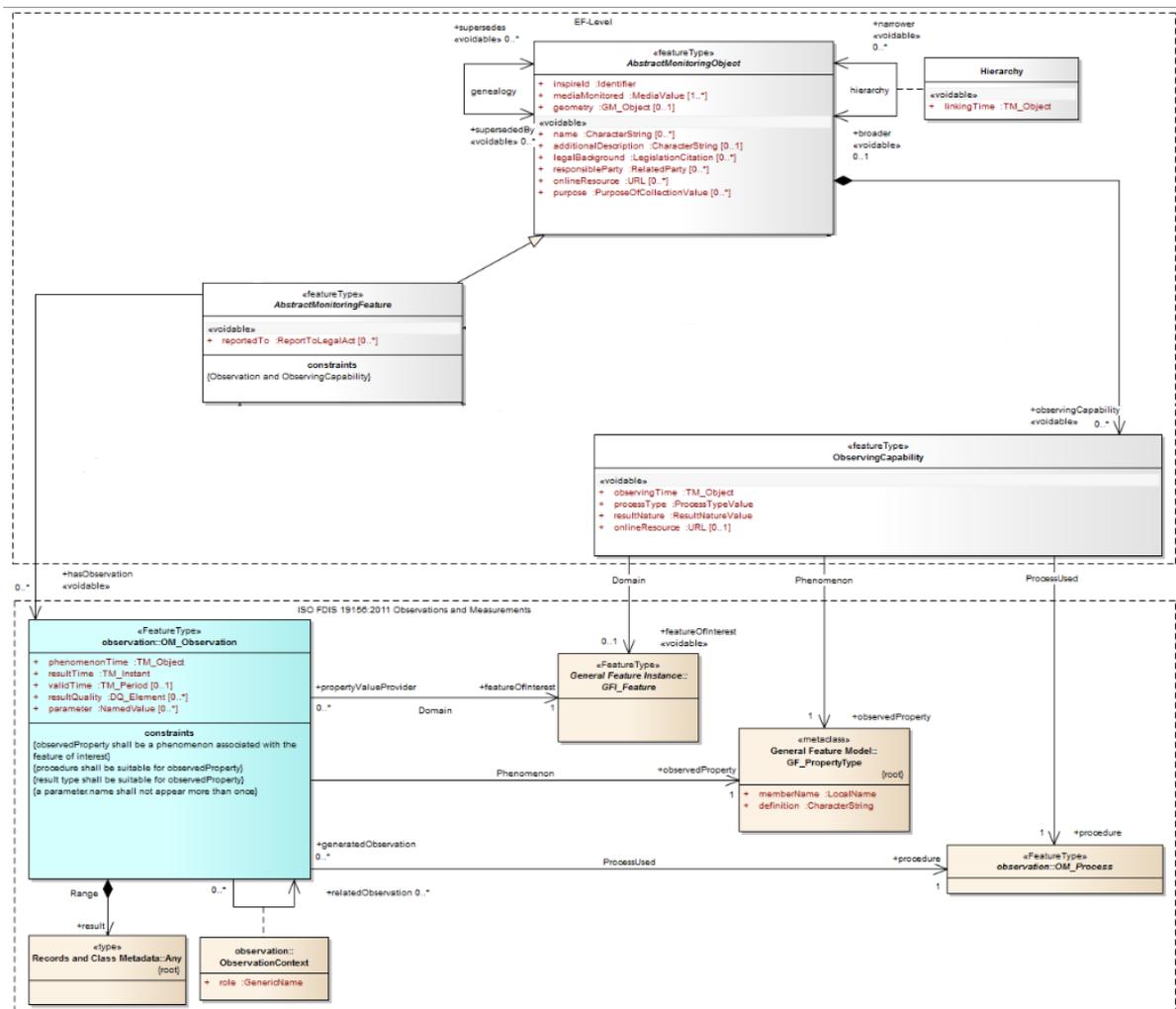


Figure 20 - Fragment of the UML Class Diagram for the “Environmental Monitoring Facilities” conceptual schema related to observations and measurements extracted from [56].

An essential part of the theme environmental monitoring facilities is to link to Observations and Measurements (O&M) [51] taken at an environmental monitoring facility.

The link to the Observation class reflects this direct connection that is possible from any environmental monitoring facility. Figure 20 shows the UML class diagram for this second aspect of the “Environmental Monitoring Facilities” application schema extracted from [56]. The concepts presented in this diagram were discussed in section 5.2.

5.4.2 Ontological Analysis of the INSPIRE Conceptual Model

In this section, we perform the ontological analysis of each UML model of INSPIRE presented previously.

Ontological Analysis of the Hydrography UML Model

In the “Hydro - Physical Waters” application schema, *HydroObject* represents hydrographic objects of the real world and is classified as a specialization of UFO-A Substantial. In addition, all of the *HydroObject* specializations (*DrainageBasin*, *SurfaceWater*, *Watercourse*, *ManMadeObject*, and *SeaArea*) are classified as specializations of UFO-A Substantial. The attributes *geographicalName* and *hydroId* refer to specializations of UFO-A Abstract Individual. Table 29 presents the relations between the “Hydro - Physical Waters” conceptual schema elements and UFO concepts.

Table 29 - Relations between the “Hydro - Physical Waters” conceptual schema elements and UFO concepts

Hydro - Physical Waters conceptual schema element	Relation Symbol	UFO concept
HydroObject	[S]	UFO-A: Substantial
DrainageBasin	[S]	UFO-A: Substantial
SurfaceWater	[S]	UFO-A: Substantial
Watercourse	[S]	UFO-A: Substantial
ManMadeObject	[S]	UFO-A: Substantial
SeaArea	[S]	UFO-A: Substantial
geographicalName	[S]	UFO-A: Abstract Individual
hydroId	[S]	UFO-A: Abstract Individual

Ontological Analysis of the Administrative Unit UML Model

Regarding the elements of the “AdministrativeUnit” spatial object, *AdministrativeUnit* represents areas or regions where a Member State has and/or exercises jurisdictional rights, for local, regional and national governance. Thus, *AdministrativeUnit* is classified as a

specialization of UFO-A Substantial. The self-reference of association or composition of the AdministrativeUnit type can also be represented in UFO by the *componentOf* relation.

The attribute *country* is a two-character country code. As a consequence, it refers to a specialization of UFO-A Abstract Individual. The attributes *geometry*, *name* and *nationalCode* refer to specializations of UFO-A Abstract Individual too. As the *nationalLevel* represents the type of the administrative unit, it refers to a specialization of UFO-A Substantial Universal. In turn, the *nationalLevelName* represents the name of the level in the national administrative hierarchy. Thus, it refers to a specialization of UFO-A Abstract Individual. Table 30 shows the relations between the “AdministrativeUnit” spatial object elements and UFO concepts.

Table 30 - Relations between the “AdministrativeUnit” spatial object elements and UFO concepts

AdministrativeUnit spatial object element	Relation Symbol	UFO concept
AdministrativeUnit	[S]	UFO-A: Substantial
country	[S]	UFO-A: Abstract Individual
geometry	[S]	UFO-A: Abstract Individual
name	[S]	UFO-A: Abstract Individual
nationalCode	[S]	UFO-A: Abstract Individual
nationalLevel	[S]	UFO-A: Substantial Universal
nationalLevelName	[S]	UFO-A: Abstract Individual

Ontological Analysis of the Environmental Monitoring Facilities UML Model

In the “Environmental Monitoring Facilities” application schema, *EnvironmentalMonitoringFacilities (EMFs)* represent objects. Thus, an EMF is classified as a specialization of UFO-A Substantial. The attribute *representativePoint* is a geographic point and refers to a specialization of UFO-A Abstract Individual. The *measurementRegime* also refers to a specialization of UFO-A Abstract Individual. The *specialisedEMFType* categorizes EMFs as platform, site, station, sensor, etc. As a consequence, it refers to a specialization of UFO-A Substantial Universal. The *operationalActivityPeriod* represents the lifespan of the physical facility and refers to a specialization of UFO-A Abstract Individual. In relation to the attributes inherited from *AbstractMonitoringObject*, *mediaMonitored* represents the type of the monitored environmental medium and refers to a specialization of UFO-A Substantial Universal. The attribute *responsibleParty* refers to a specialization of UFO-C Agent.

An *EnvironmentalMonitoringProgramme (EMP)* is classified as a specialization of UFO-C Normative Description. The attribute *geometry* of EMP represents a geometric area and refers to a specialization of UFO-A Abstract Individual. The *legalBackground* represents a normative and refers to a specialization of UFO-C Normative Description.

The *EnvironmentalMonitoringActivity (EMA)* is an event and is classified as a specialization of UFO-B Event. The attribute *activityTime* represents the lifespan of the EMA. It refers to specializations of UFO-B Event begin and end Time Points. The *responsibleParty* of EMA refers to a specialization of UFO-C Agent. The association between EMA and EMFs (also between EMA and EMPs) refers to the participation of substantials in events of UFO-C. Table 31 presents the relations between the “Environmental Monitoring Facilities” conceptual schema elements and UFO concepts.

Table 31 - Relations between the “Environmental Monitoring Facilities” conceptual schema elements and UFO concepts

Environmental Monitoring Facilities conceptual schema element	Relation Symbol	UFO concept
EnvironmentalMonitoringFacility	[S]	UFO-A: Substantial
representativePoint	[S]	UFO-A: Abstract Individual
measurementRegime	[S]	UFO-A: Abstract Individual
specialisedEMFType	[S]	UFO-A: Substantial Universal
operationalActivityPeriod	[S]	UFO-A: Abstract Individual
mediaMonitored	[S]	UFO-A: Substantial Universal
responsibleParty	[S]	UFO-C: Agent
EnvironmentalMonitoringProgramme	[S]	UFO-C: Normative Description
legalBackground	[S]	UFO-C: Normative Description
EnvironmentalMonitoringActivity	[S]	UFO-B: Event
activityTime	[S]	UFO-B: begin and end Time Points

5.5 The ISO/TC 211

The ISO/TC 211 [57] is concerned with the standardization in the field of digital geographic information. It establishes a structured set of standards for information concerning objects or phenomena that are directly or indirectly associated with a location relative to the Earth. These standards may specify, for geographic information, methods, tools and services for data management (including definition and description), acquiring, processing, analyzing, accessing, presenting and transferring such data in digital/electronic form between different users, systems and locations. In section 5.2, we have discussed the ISO 19156:2011 (O&M)

[51], which is part of the scope of ISO/TC 211. In this section, we present some concepts of the ISO 19111:2007 Referencing by Coordinates Standard [58].

5.5.1 Overview of the Coordinate Reference System UML Schema

The ISO 19111:2007 defines the conceptual schema for the description of referencing by coordinates [58]. Figure 21 shows the UML class diagram for the Coordinate Reference System package extracted from [58].

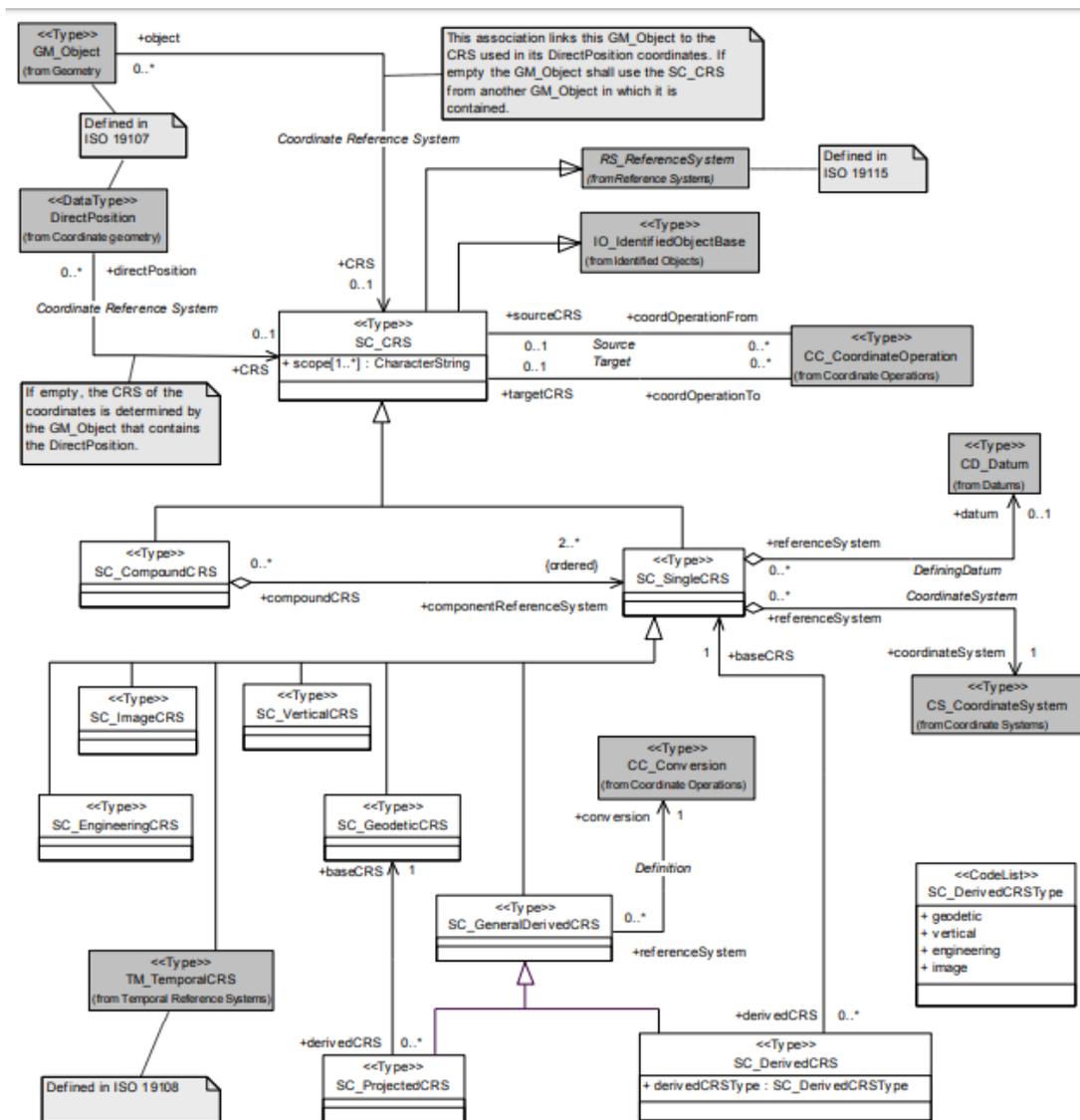


Figure 21 - UML Class Diagram for the Coordinate Reference System package extracted from [58].

A *coordinate* is one of n scalar values that define the position of a single point in n-dimensional space. A *coordinate tuple* is an ordered list of n coordinates that define the position of a single point in n-dimensional space. The number of coordinates is equal to the dimension of the coordinate space [58].

5.5.2 Ontological Analysis of the Coordinate Reference System UML Schema

In the Coordinate Reference System package, *coordinate* is a value and *coordinate tuple* is an ordered sequence of values. Thus, they are classified as specializations of UFO-A Basic and Composed Measurement Reference Region, respectively. *Coordinate System Axes* are the dimensions of the coordinate space and are classified as specializations of UFO-A Measurement Reference Dimension. *Coordinate System* is composed of a non-repeating sequence of coordinate system axes. As a result, it is classified as a specialization of UFO-A Measurement Reference Domain. *Datum* is a set of parameters that defines the position of the origin, the scale, and the orientation of a coordinate system and is classified as a specialization of UFO-C Normative Description. *Coordinate Reference System* is also a specialization of UFO-C Normative Description that includes a datum and defines a UFO-A Quality Function to measure location. Finally, a *unit of measurement* is a defined quantity in which dimensioned parameters are expressed. It is characterized by the parameter that is being measured and the value associated to this parameter. The first refers to a particular UFO-A Quality Universal (location). The second refers to a UFO-A Unit (a specialization of UFO-A Measurement Quality Region). Table 32 summarizes the relations between the Coordinate Reference System UML schema elements and UFO concepts.

Table 32 - Relations between the Coordinate Reference System UML schema elements and UFO concepts

Coordinate Reference System UML schema element	Relation Symbol	UFO concept
coordinate	[S]	UFO-A: Basic Measurement Reference Region
coordinate tuple	[S]	UFO-A: Composed Measurement Reference Region
Coordinate System Axis	[S]	UFO-A: Measurement Reference Dimension
Coordinate System	[S]	UFO-A: Measurement Reference Domain
Datum	[S]	UFO-C: Normative Description
Coordinate Reference System	[S]	UFO-C: Normative Description
unit of measurement (parameter)	[E]	UFO-A: Quality Universal (a "location" quality universal)
unit of measurement (value)		UFO-A: Unit (a specialized Measurement Quality Region)

5.6 The Environment Ontology (EnvO)

The Environment Ontology (EnvO) [59][60] provides a controlled, structured vocabulary that is designed to support the annotation of any organism or biological sample with environment descriptors. It is grounded in the Basic Formal Ontology (BFO) [48] and is available in OWL

and OBO formats. EnvO contains terms for biomes (e.g., tropical rain forest biome), environmental features (e.g., mountain, pond), and environmental material (e.g., sediment, soil, water, and air). These three sets of terms enable a concise, standardized, and comprehensive description of environment that is key to the integration, archiving and federated searching of environmental data. In this work, we are interested in the environmental material terms.

5.6.1 Overview of the EnvO Material Terms

Figure 23 shows a tree view of part of the EnvO related to material terms extracted from [61]. A *continuant* is an entity that persists, endures, or continues to exist through time while maintaining its identity. An *independent continuant* is a continuant entity that is the bearer of qualities (e.g., an organism, a spatial region). A *material entity* is an independent continuant that has some portion of matter as proper or improper continuant part (e.g., a human being, the undetached arm of a human being, an aggregate of human beings). *Fiat object parts* are material entities distinguished by fiat within larger object wholes (e.g., mountains demarcated within mountain ranges) [61]. These are BFO concepts reused by EnvO.

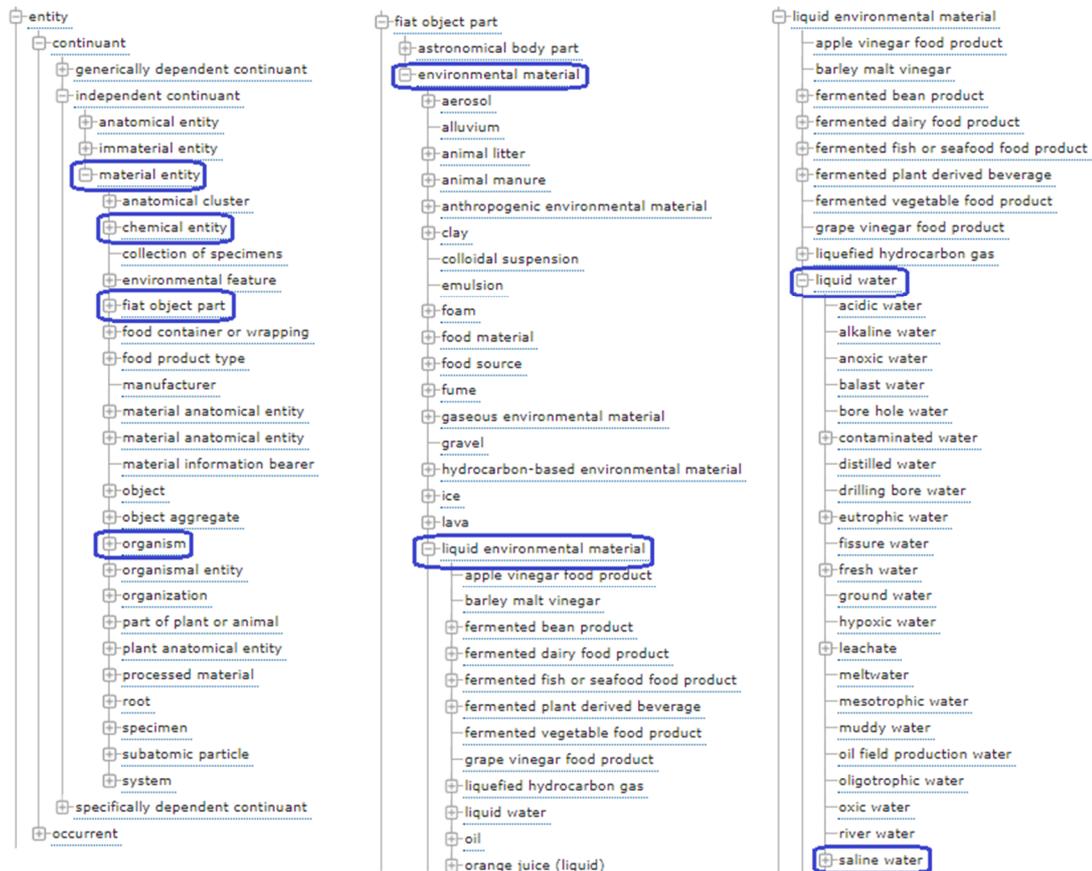


Figure 23 - Tree view of part of the EnvO related to material terms [61].

A portion of *environmental material* is a fiat object which forms the medium or part of the medium of an environmental system. A *liquid environmental material* is an environmental material that is in a liquid state. *Liquid water* is an environmental material primarily composed of dihydrogen oxide in its liquid form. *Saline water* is a water that contains a significant concentration of dissolved salts, and so on [61].

Sediment is an environmental material comprised of any particulate matter that can be transported by fluid flow and which eventually is deposited as a layer of solid particles on the bed or bottom of a body of water or other liquid [61].

An *organism* is a material entity that is an individual living system, such as animal, plant, bacteria or virus, which is capable of replicating or reproducing, growth and maintenance in the right environment. An organism may be unicellular or made up, like humans, of many billions of cells divided into specialized tissues and organs [61].

A *chemical entity* is a physical entity of interest in chemistry including molecular entities, parts thereof, and chemical substances [61]. It is a ChEBI concept reused by EnvO. Many other concepts are defined, but these are sufficient for understanding the material terms of EnvO that apply to that work.

5.6.2 Ontological Analysis of the EnvO Material Terms

As explained before, EnvO is grounded in BFO. The BFO concept *continuant* is equivalent to the UFO-A Endurant. *Independent continuant*, *material entity* and *fiat object parts* are classified as specializations of UFO-A Substantial. The EnvO elements *environmental material*, *liquid environmental material*, *liquid water*, *saline water*, *sediment*, *organism* and *chemical entity* are classified as specializations of UFO-A Substantial too. Table 33 summarizes the relations between the EnvO Material Terms elements and UFO concepts.

Table 33 - Relations between the EnvO Material Terms elements and UFO concepts

EnvO Material Terms element	Relation Symbol	UFO concept
continuant	[E]	UFO-A: Endurant
independent continuant	[S]	UFO-A: Substantial
material entity	[S]	UFO-A: Substantial
fiat object part	[S]	UFO-A: Substantial
environmental material	[S]	UFO-A: Substantial
liquid environmental material	[S]	UFO-A: Substantial
liquid water	[S]	UFO-A: Substantial
saline water	[S]	UFO-A: Substantial
sediment	[S]	UFO-A: Substantial
organism	[S]	UFO-A: Substantial
chemical entity	[S]	UFO-A: Substantial

5.7 The ChEBI Ontology

The Chemical Entities of Biological Interest (ChEBI) [62] is a freely available dictionary of molecular entities focused on “small” chemical compounds. The term “molecular entity” encompasses any constitutionally or isotopically distinct atom, molecule, ion, ion pair, radical, radical ion, complex, conformer, etc., identifiable as a separately distinguishable entity. The molecular entities in question are either products of nature or synthetic products used to intervene in the processes of living organisms (either deliberately, as for drugs, or unintentionally, as for chemicals in the environment). The qualifier “small” implies the exclusion of entities directly encoded by the genome, and thus as a rule nucleic acids, proteins and peptides derived from proteins by cleavage are not included. Classes of molecular entities and part-molecular entities (in the form of substituent groups or atoms) are also included in ChEBI.

In addition ChEBI incorporates an ontology, whereby the relationships between compounds, groups or classes of compounds and their parents, children and/or siblings are specified. Its structure is essentially that of a directed acyclic graph, which differs from a simple taxonomy in that a child term can have many parent terms. Additionally, a number of relationships are incorporated which are cyclic in nature [62].

The ChEBI Ontology [62] is subdivided into three separate subontologies:

- Molecular structure, in which molecular entities or parts thereof are classified according to composition and structure, e.g., hydrocarbons, carboxylic acids, tertiary amines;
- Role, divided into three sub-categories: “chemical role” that classifies entities on the basis of their role within a chemical context, e.g., as ligand, inhibitor, surfactant; “biological role” that classifies entities on the basis of their role within a biological context, e.g., antibiotic, antiviral agent, coenzyme, hormone; and “application” that classifies on the basis of their intended use by humans, e.g., pesticide, antirheumatic drug, fuel;
- Subatomic Particle, which classifies particles that are smaller than atoms, e.g., electron, photon, nucleon.

This ontology is provided in OWL and OBO formats. In this work, we are interested in the *Molecular structure* ontology.

5.7.1 Overview of the ChEBI Molecular Structure Ontology

Figure 24 shows a tree view of part of the ChEBI Molecular Structure Ontology extracted from [63]. On the left side, the class “chemical entity” and its direct subclasses (“chemical substance”, “molecular entity”, “group” and “atom”) are shown. On the right side, different classifications of the molecular entity “calcium carbonate” are shown.



Figure 24 - Tree view of part of the ChEBI Molecular Structure Ontology extracted from [63].

According to [63], *chemical entity* is “a physical entity of interest in chemistry including chemical substances, molecular entities and parts thereof”. A *chemical substance* is “a portion of matter of constant composition, composed of molecular entities of the same type or of different types”. A *molecular entity* is “any constitutionally or isotopically distinct atom, molecule, ion, ion pair, radical, radical ion, complex, conformer etc., identifiable as a separately distinguishable entity”. A *group* is “a defined linked collection of atoms or a single atom within a molecular entity”. An *atom* is “a chemical entity constituting the smallest component of an element having the chemical properties of the element”. A *calcium carbonate* is “a calcium salt (a molecular entity) with formula CCaO_3 ”. Many other classes are defined, but these are sufficient for understanding the concepts of ChEBI Molecular Structure Ontology that apply to our work.

5.7.2 Ontological Analysis of the ChEBI Molecular Structure Ontology

Given the previous definitions of the ChEBI Molecular Structure Ontology elements and taking into account the alignment of ChEBI with BFO (where chemical entity is a specialization of BFO material entity, see Figure 23), we can take as instances of chemical

entity, for example, an individual cluster of atoms under a scanning electron microscope tip, a sodium ion in glass of brine, etc. Therefore, *chemical entity*, *chemical substance*, *molecular entity*, *group*, *atom* and *calcium carbonate* are classified as specializations of UFO-A Substantial. Table 34 summarizes the relations between the ChEBI Molecular Structure Ontology elements and UFO concepts.

Table 34 - Relations between the ChEBI Molecular Structure Ontology elements and UFO concepts

ChEBI Molecular Structure Ontology element	Relation Symbol	UFO concept
chemical entity	[S]	UFO-A: Substantial
chemical substance	[S]	UFO-A: Substantial
molecular entity	[S]	UFO-A: Substantial
group	[S]	UFO-A: Substantial
atom	[S]	UFO-A: Substantial
calcium carbonate	[S]	UFO-A: Substantial

5.8 Concluding Remarks

In this chapter, we have presented the knowledge resources selected for reused in the development of the network of reference ontologies for the integration of water quality data. We have analyzed these knowledge resources in the light of UFO, checking the relations between the knowledge resources elements and UFO concepts. As a result, this analysis provides the classification of the knowledge resources elements according to UFO categories. This makes it possible to integrate these elements into the ontology network proposed in the next chapter, since this ontology network is grounded in UFO.

6 The Network of Reference Ontologies for the Integration of Water Quality Data

In this chapter, we design and evaluate the network of reference ontologies for the integration of water quality data from the Doce River Basin. For that, different concepts related to the water quality domain have been represented: the research activities performed to produce environmental data (e.g., sampling, sample preparation, measurement, etc.); the methods and the devices used to perform these activities; the actors involved; the water quality monitoring sites; the material entities (e.g., water, sediment and aquatic biota) analyzed for the verification of the water quality of a given site; the water quality properties checked (physical, chemical and biological properties); among others.

Most of these concepts (42 out of a total of 78 concepts, i.e., 53.8%) were reused from the knowledge resources selected for reuse with the application of CLeAR to the water quality domain (INSPIRE, O&M, ISO/TC 2011, QUDT, EnvO and ChEBI). New 36 concepts have been added as needed. This aims to promote reuse and avoid unnecessary proliferation of new ontologies. We aim to ensure that alignment with the knowledge resources that were selected in the application of CLeAR is possible. Therefore, we indicate the relations between the proposed ontology network concepts and the elements of the existing knowledge resources through traceability tables (i.e., tables that indicate the provenance of the ontology network concepts).

As explained earlier, due to the complexity and the characteristics of the water quality domain, the ontology network was organized in the layered architecture proposed by [28] and adopts the Unified Foundational Ontology (UFO) [24][25][26] at the foundational level to ground core and domain level ontologies.

This chapter is structured as follows. Section 6.1 presents the ontology network development process. Section 6.2 presents the ontology network architecture. Section 6.3 addresses the core level ontologies. Section 6.4 addresses the domain level ontologies. Section 6.5 evaluates the ontology network. Section 6.6 discusses related work. Finally, section 6.7 presents concluding remarks.

6.1 The Ontology Network Development Process

To develop the proposed ontology network, we adopted some NeOn methodology [10] guidelines in combination with the guidelines proposed by CLeAR. From applying CLeAR to the water quality domain, we define the ontology network requirements (CLeAR cycle I), identify existing knowledge resources about the water quality domain (CLeAR cycle II), and select the knowledge resources to be reused in the development of the ontology network (CLeAR cycle III). These activities correspond to NeOn's specification of ontology requirements, search for reusable knowledge resources, assessment of candidate knowledge resources, and selection of knowledge resources. The main products of them are: integration questions, data sources to be integrated, domain aspects, existing set of knowledge resources about the water quality domain, and knowledge resources selected for reuse in the construction of the ontology network.

As the knowledge resources selected for reuse differ from each other and cannot be integrated into their original format, we performed an ontological analysis of them based on the Unified Foundational Ontology (UFO) [24][25][26]. This analysis reveals the correspondences between the knowledge resources elements and UFO concepts. This makes it possible to adjust previous knowledge resources or portions of them for integration into the ontology network. This activity corresponds to NeOn's adaptation of selected knowledge resources. All knowledge resources elements needed to represent the domain aspects or the elements of data sources to be integrated, or needed to answer the integration questions have been aligned with UFO concepts to be reused in the construction of the ontology network.

Then we performed NeOn's ontology conceptualization activity, in which the network of reference ontologies was modeled according to the layered architecture proposed by [28]. In this activity, the knowledge resources elements that represent domain aspects related to research activities, spatial location and material entities were reused to build core ontologies. They have been included in the ontology network as specializations of UFO concepts (the same applies to their relationships). In addition, the knowledge resources elements that represent domain aspects related to environmental monitoring and water quality were reused to build domain ontologies. They were included in the ontology network as specializations of core or UFO concepts (the same applies to their relationships). New concepts have also been added to the ontology network through specializations of UFO or core concepts as needed.

NeOn's ontology formalization and ontology implementation activities were not performed because we did not build an operational version of the ontology network in this work.

Finally, we performed NeOn's ontology evaluation by verifying and validating the ontology network. In the verification activity proposed by NeOn, one should check whether the modeled elements answer the competence questions. All modeled elements must be used to answer the CQs. Due to the characteristics of this work, we performed this activity by verifying integration questions rather than competency questions. As we have used a non-exhaustive list of IQs, only the ontology network elements needed to answer them were covered. In turn, the validation activity can be performed through expert judgment or ontology instantiation. As we need to articulate data semantics, we instantiated the ontology network with water quality data provided by different sources. Besides that, we show how the ontology network elements can be used to annotate such data.

In the previous chapters, we presented the application of the CLeAR approach to the water quality domain and the ontological analysis of the knowledge resources selected for reuse in the construction of the ontology network. Next, we present the network of reference ontologies for the integration of water quality data and the ontology network evaluation.

6.2 The Ontology Network Architecture

Figure 25 presents the current ontology network architecture. At the foundational level, there is the *Unified Foundational Ontology (UFO)* [24][25][26]. UFO concepts are used to ground the ontologies of the core and domain levels. At the core level, there are three core ontologies: *Material Entity Ontology*, *Spatial Location Ontology* and *Scientific Research Activity Ontology* (divided into subontologies *Research Activity Ontology*, *Sampling Ontology*, *Preparation Ontology* and *Measurement Ontology*). They form the basis for domain level ontologies. At the domain level, there are two ontologies: *Water Quality Ontology* and *Environmental Monitoring Ontology*. Following, core and domain level ontologies are presented.

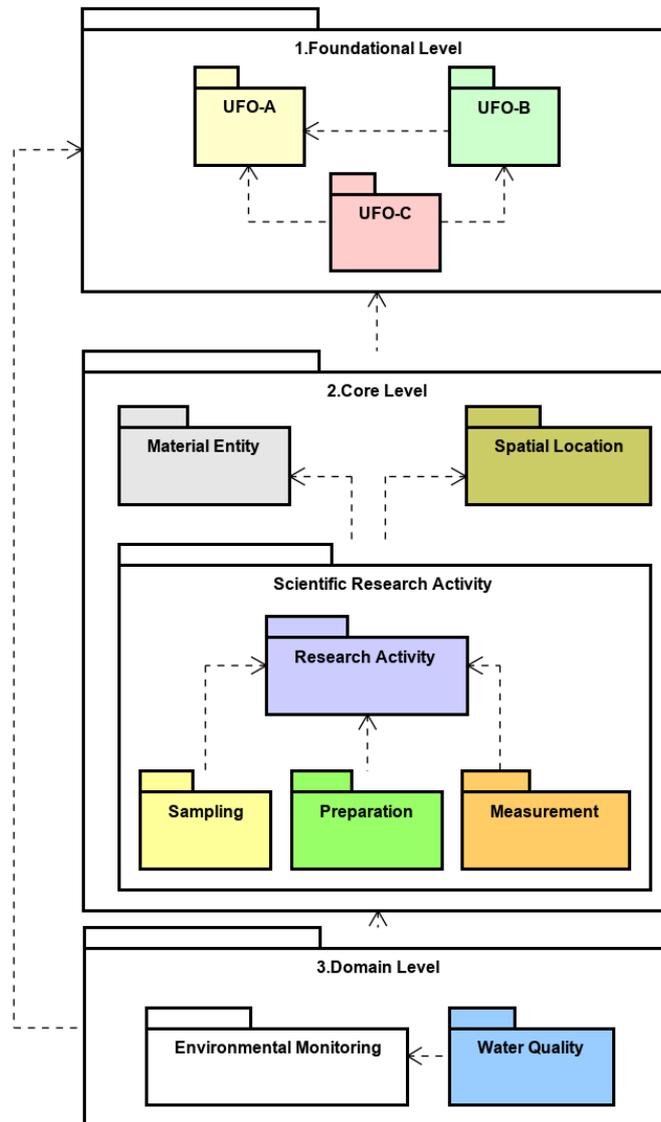


Figure 25 - Architecture of the Network of Reference Ontologies for the Integration of Water Quality Data.

6.3 The Core Level Ontologies

The core level ontologies of the proposed ontology network provide knowledge about material entities, spatial location and scientific research activities. This knowledge is common to the different subdomains of the environmental domain (e.g., water quality, air quality, observation of the taxon of an animal, etc.). Thus, they must be modeled at the core level to be reused by subdomains. As mentioned before, there are three core ontologies: *Material Entity Ontology*, *Spatial Location Ontology* and *Scientific Research Activity Ontology* (divided into subontologies *Research Activity Ontology*, *Sampling Ontology*, *Preparation Ontology* and *Measurement Ontology*). Figure 26 shows an integrated view of them. Next, they are detailed.

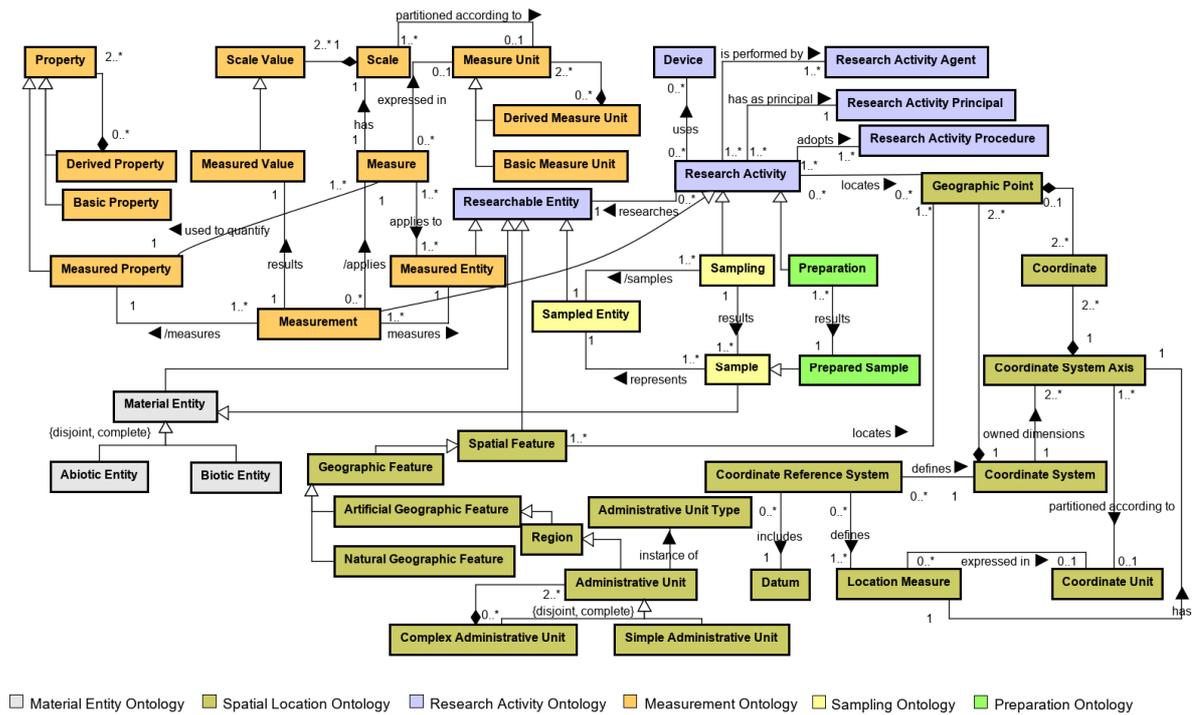


Figure 26 - The Core Level Ontologies.

6.3.1 The Material Entity Ontology

The *Material Entity Ontology* comprises concepts for dealing with the existing types of material entities (see Figure 27). It was developed based on the EnvO Material Terms. The main concept is *Material Entity*, a specialization of UFO-A Substantial. Material Entity specializes in *Abiotic Entity* (non-living parts of an environment such as water, air, soil, etc.) and *Biotic Entity* (living parts of an environment such as animals, plants, etc.).

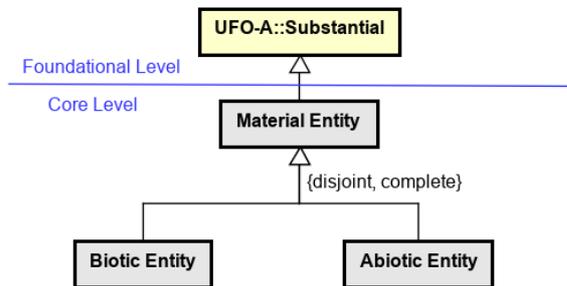


Figure 27 - The Material Entity Ontology.

Table 35 presents the EnvO Material Terms elements whose adapted reuse resulted in each *Material Entity Ontology* concept.

Table 35 - Correspondences between Material Entity Ontology concepts and EnvO Material Terms elements

Material Entity Ontology concept	EnvO Material Terms element
Material Entity	Material Entity
Biotic Entity (new concept)	-
Abiotic Entity (new concept)	-

6.3.2 The Spatial Location Ontology

Figure 28 presents the *Spatial Location Ontology*. This ontology provides concepts related to spatial features, i.e., anything with spatial extant (size, shape, or position) [64], and geographic points. Examples of spatial features are: a city, a country, a mountain, a river, etc.

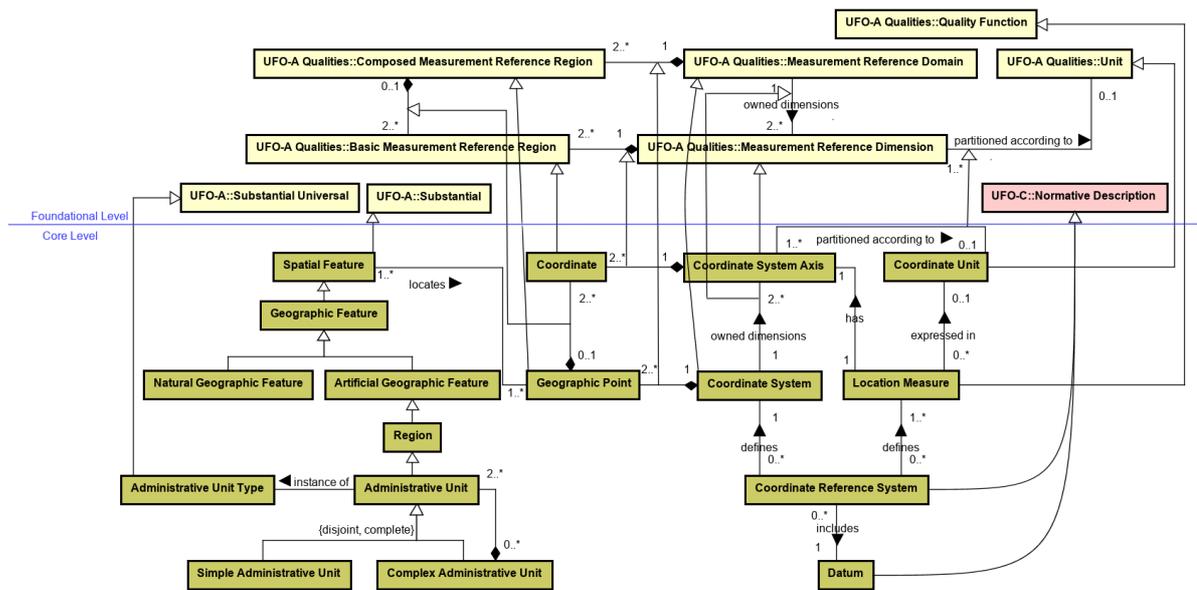


Figure 28 - The Spatial Location Ontology.

Spatial Feature is a specialization of UFO-A Substantial. *Spatial Feature* is specialized in *Geographic Feature*. Geographic features are naturally-created (e.g., a river, a mountain) or artificially-created spatial features (e.g., a city, a dam, a water treatment plant). *Artificial Geographic Feature* is specialized in *Region*. *Regions* are areas that have definable characteristics (e.g., common geographic features, language or government) but not always fixed boundaries. Some examples of regions are: a country, a city and the South Atlantic Hydrographic Region. *Region* specializes in *Administrative Unit* (e.g., a country, a city).

Administrative Unit is an instance of *Administrative Unit Type* (e.g., country and city) that is a specialization of UFO-A Substantial Universal. Administrative units can be simple or complex. In the last case, they are composed of two or more administrative units (e.g., a country, a state). Finally, spatial features are located in geographic points.

Geographic Point represents a coordinate tuple of a spatial location in a given coordinate system and is a specialization of UFO-A Composed Measurement Reference Region. Geographic Point is composed of two or more *Coordinates* that are specializations of UFO-A Basic Measurement Reference Region. *Coordinate System Axes* are the dimensions of the coordinate space and are specializations of UFO-A Measurement Reference Dimension. *Coordinate System* is composed of a non-repeating sequence of coordinate system axes and is a specialization of UFO-A Measurement Reference Domain. *Datum* is a set of parameters that defines the position of the origin, the scale, and the orientation of a coordinate system and is classified as a specialization of UFO-C Normative Description. *Coordinate Reference System* is also a specialization of UFO-C Normative Description that includes a datum and defines a UFO-A Quality Function to measure location (*Location Measure*). Finally, Location Measure can be expressed in *Coordinate Units* (e.g., decimal degrees). These units partition the coordinate system axes and are specializations of UFO-A Units.

Concepts related to administrative units were modeled based on the Administrative Unit UML Model of INSPIRE. Concepts related to geographic points were modeled based on the Coordinate Reference System UML Schema (ISO/TC 211). Other concepts were created to complement the *Spatial Location Ontology*. Although we have borrowed the description of spatial feature of “Spatial Thing” from [64], the definition of “Feature” [65] is a better semantic fit for spatial feature as it is explicitly specified as being disjoint from geometry.

Table 36 presents the knowledge resources elements whose adapted reuse resulted in each *Spatial Location Ontology* concept.

Table 36 - Correspondences between Spatial Location Ontology concepts and knowledge resources reused elements

Spatial Location Ontology concept	Knowledge Resource reused	Knowledge Resource reused element
Spatial Feature (new concept)	-	-
Geographic Feature (new concept)	-	-
Natural Geographic Feature (new concept)	-	-
Artificial Geographic Feature (new concept)	-	-
Region (new concept)	-	-
Administrative Unit	Administrative Unit UML Model of INSPIRE	AdministrativeUnit
Simple Administrative Unit (new concept)	-	-
Complex Administrative Unit (new concept)	-	-
Administrative Unit Type	Administrative Unit UML Model of INSPIRE	nationalLevel
Geographic Point	Coordinate Reference System UML Schema (ISO/TC 211)	coordinate tuple
Coordinate	Coordinate Reference System UML Schema (ISO/TC 211)	coordinate
Coordinate System Axis	Coordinate Reference System UML Schema (ISO/TC 211)	Coordinate System Axis
Coordinate System	Coordinate Reference System UML Schema (ISO/TC 211)	Coordinate System
Datum	Coordinate Reference System UML Schema (ISO/TC 211)	Datum
Coordinate Reference System	Coordinate Reference System UML Schema (ISO/TC 211)	Coordinate Reference System
Coordinate Unit	Coordinate Reference System UML Schema (ISO/TC 211)	unit of measurement
Location Measure (new concept)	-	-

6.3.3 The Scientific Research Activity Ontology

The *Scientific Research Activity Ontology* deals with the different types of research activities performed in empirical research, such as (physical) sampling, sample preparation, measurement, etc. It was developed based on some concepts of the O&M Conceptual Model and QUDT Ontologies.

Regarding research activities, we have identified that some characteristics are common to all types of research activities, such as temporal and spatial properties, actors involved in their execution, responsible actors, among others. They are related to provenance information and are generally addressed by *metadata*, but the modeling of research activity shows that

they are properties of events. Thus, we have created a subontology to represent these properties: the *Research Activity Ontology*.

This subontology may be specialized to handle the intrinsic characteristics of each type of research activity. We have specialized it in the following subontologies: *Sampling Ontology*, *Preparation Ontology* and *Measurement Ontology*. However, new specializations can be made to deal with other types of research activities, such as observations (e.g., an observation of the taxon of a beetle), assays, among others. Figure 29 shows the *Scientific Research Activity Ontology*. Following, their subontologies are described.

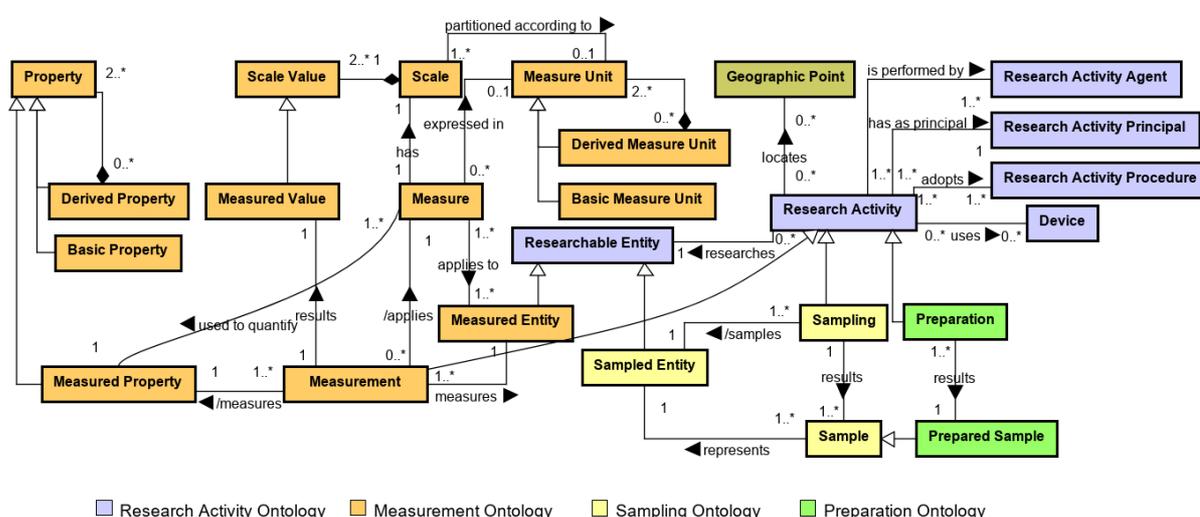


Figure 29 - The Scientific Research Activity Ontology.

The Research Activity Ontology

The *Research Activity Ontology* comprises concepts that are common to the different types of research activities (see Figure 30). *Research Activity* is a specialization of UFO-B Event used to generalize these types. Research activities are characterized by temporal and spatial properties, as well as the researched entity. Regarding temporal properties, research activities inherit begin and end Time Points from UFO-B. In relation to spatial properties, Geographic Point represents the coordinates tuple corresponding to the spatial location of a research activity. *Researchable Entity* is a specialization of UFO-A Individual because it can be a substantial (e.g., a river, a city) or an event (such as a process). A research activity is also characterized by the procedure adopted and the device employed. *Research Activity Procedure* is a specialization of UFO-C Normative Description that defines the rules to be followed for the execution of a research activity. *Device* is a specialization of UFO-C Physical

Object. Examples of devices are: collectors, sensors, etc. In order to capture provenance, the *Agents* involved in the execution and the agent responsible for a research activity (the so-called *Principal*) are identified. They are specializations of UFO-C Agent and can be physical (such as researches) or social agents (governmental agencies, research institutions, laboratories, etc.).

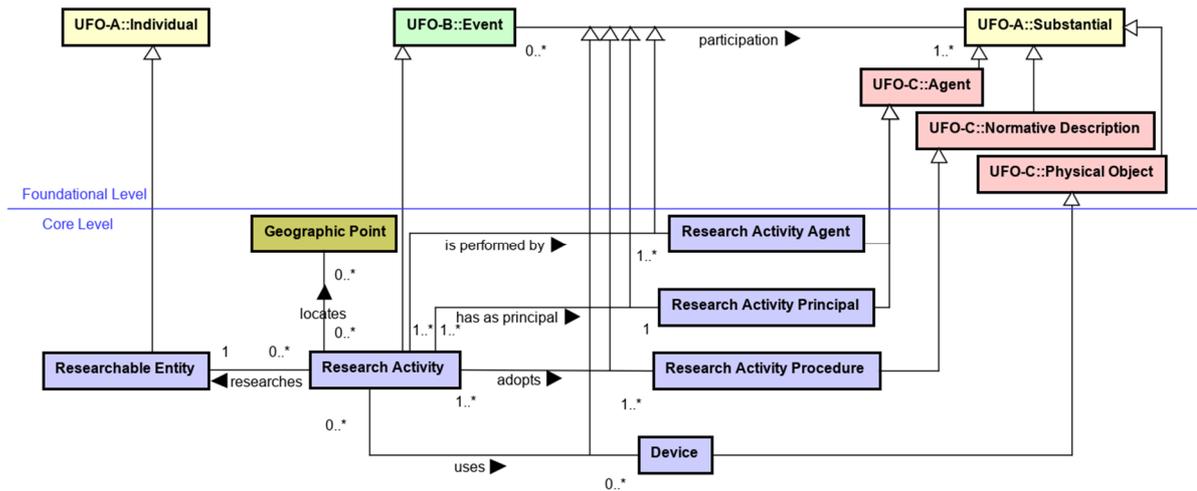


Figure 30 - The Research Activity Ontology.

Table 37 presents the O&M Conceptual Model elements whose adapted reuse resulted in each *Research Activity Ontology* concept.

Table 37 - Correspondences between Research Activity Ontology concepts and O&M Conceptual Model elements

Research Activity Ontology concept	O&M Conceptual Model element
Research Activity (new concept)	-
Researchable Entity	Feature (feature of interest)
Research Activity Agent	processOperator
Research Activity Principal (new concept)	-
Research Activity Procedure	procedure (method), samplingMethod
Device	procedure (instrument)

The Sampling Ontology

The *Sampling Ontology*, presented in Figure 31, deals with concepts related to the sampling activity. *Sampling* is the collection of samples for in situ and/or laboratory analysis. Sampling is a specialization of Research Activity, inheriting concepts related to research activity. *Sampled Entity* is a specialization of Researchable Entity and represents the target research entity. *Sample* represents a portion of a sampled entity that must be studied with the ultimate

goal of characterizing the sampled entity. Sample is a specialization of UFO-A Substantial. For instance, in the case of a water quality research of a river, a sample of water or sediment can be collected to verify the river water quality.

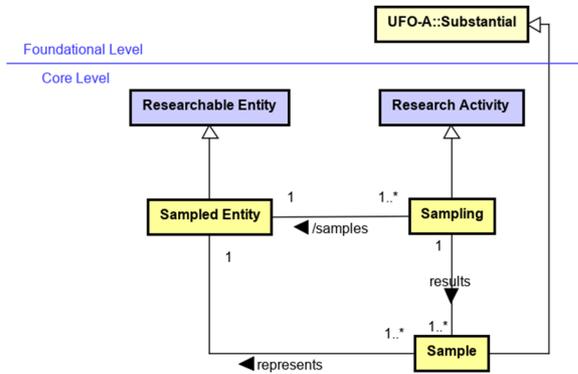


Figure 31 - The Sampling Ontology.

Table 38 presents the O&M Conceptual Model elements whose reuse resulted in each *Sampling Ontology* concept.

Table 38 - Correspondences between Sampling Ontology concepts and O&M Conceptual Model elements

Sampling Ontology concept	O&M Conceptual Model element
Sampling (new concept)	-
Sampled Entity	Feature (sampled feature)
Sample	Specimen

The Preparation Ontology

The *Preparation Ontology*, shown in Figure 32, addresses concepts related to the sample preparation activity. It refers to the ways in which a sample is treated before being analyzed. *Preparation* is a specialization of *Research Activity*. *Prepared Sample* represents a sample that has been prepared for measurement. Not all samples need to be prepared before they are measured.

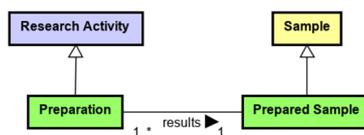


Figure 32 - The Preparation Ontology.

Table 39 presents the O&M Conceptual Model elements whose reuse resulted in each *Preparation Ontology* concept.

Table 39 - Correspondences between Preparation Ontology concepts and O&M Conceptual Model elements

Preparation Ontology concept	O&M Conceptual Model element
Preparation	PreparationStep
Prepared Sample (new concept)	-

The Measurement Ontology

The *Measurement Ontology* (see Figure 33) provides concepts related to the measurement activity. Most of the concepts presented here were extracted from the Core Ontology on Measurement (COM) presented in [50]. COM was not returned by the application of CLeAR to the water quality domain because it does not address the environmental domain, but only concepts related to the measurement aspect. However, in addition to cover most of the O&M Conceptual Model and QUDT Ontologies concepts selected for reuse in the construction of the *Measurement Ontology*, it was developed in alignment with UFO. For this reason, we have reused its concepts in this work.

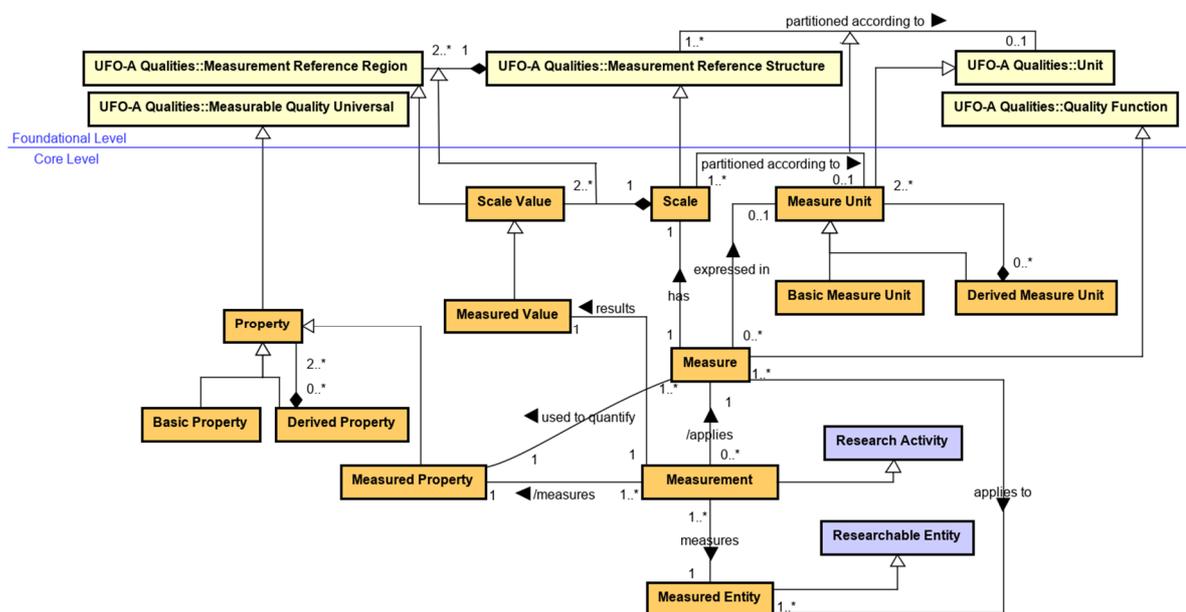


Figure 33 - The Measurement Ontology.

Measurement can be defined as a set of actions aiming to characterize an entity by attributing values to its properties. *Measurement* is a specialization of *Research Activity*. *Measured Entity* is a specialization of *Researchable Entity*. It represents an entity that has one

or more measured properties, such as a person, a water sample, etc. *Property* is a specialization of UFO-A Measurable Quality Universal that deals with qualities of entities. It specializes in basic and derived property. *Basic Property* is a specialization of UFO-A Simple Quality Universal that does not depend on other properties to be measured (e.g., weight and height). *Derived Property* is a specialization of UFO-A Composed Quality Universal that depends on others to be measured (for example, Body-Mass Index). *Measured Property* represents a property that is measured. *Measures* are used for quantifying measured properties. Measure is a specialization of UFO-A Quality Function in the sense that it maps an instance of measured property to a measured value. Measures have *Scales* composed by all possible values (*Scale Value*) to be associated to a measured property. Scale is a specialization of UFO-A Measurement Reference Structure and Scale Value is a specialization of UFO-A Measurement Reference Region. Measures can be expressed in *Units* (e.g., meter, kilogram). A measure unit in which a measure is expressed partitions its scale. For instance, if the measure height is expressed in meters, it means that its scale (a linear structure isomorphic to the positive half-line of the real numbers) is partitioned in meters. Note that the UFO fragment used to ground the *Measurement Ontology* is the same as that used to ground the portion of the *Spatial Location Ontology* related to geographic points and coordinates. This is because this portion of the *Spatial Location Ontology* address the measurement of spatial location.

Table 40 presents the knowledge resources elements whose adapted reuse resulted in each *Measurement Ontology* concept.

Table 40 - Correspondences between Measurement Ontology concepts and knowledge resources reused elements

Measurement concept	Knowledge Resource reused	Knowledge Resource reused element
Measurement	O&M Conceptual Model	Measurement
Measured Entity	O&M Conceptual Model	Feature (feature of interest, sampled feature), Specimen
Measured Property	QUDT Ontologies	Quantity Kind
Property	QUDT Ontologies	Quantity Kind
Basic Property	QUDT Ontologies	Base Quantity Kind
Derived Property	QUDT Ontologies	Derived Quantity Kind
Measure (new concept)	-	-
Scale (new concept)	-	-
Scale Value (new concept)	-	-
Measured Value	QUDT Ontologies	Quantity Value
Measure Unit	QUDT Ontologies	Unit
Measure Basic Unit	QUDT Ontologies	Base Unit
Measure Derived Unit	QUDT Ontologies	Derived Unit

6.4 The Domain Level Ontologies

The domain level ontologies of the proposed ontology network provide knowledge about environmental monitoring and water quality domain. Knowledge related to environmental monitoring is specific to some environmental subdomains (e.g., water quality, air quality). It does not extend to all environmental subdomains (e.g., observation of the taxon of an animal). So it was modeled at this level. There are two domain ontologies: *Water Quality Ontology* and *Environmental Monitoring Ontology*. Next, they are detailed.

6.4.1 The Environmental Monitoring Ontology

The *Environmental Monitoring Ontology* defines concepts related to environmental monitoring, monitoring points, monitoring programs and monitoring facilities (see Figure 34). It was modeled based on the Environmental Monitoring Facilities UML Model of INSPIRE. *Monitoring* consists of a set of research activities, performed periodically, for environmental quality control. *Monitoring* is a specialization of UFO-B Complex Event because it is composed of other research activities, such as sampling and measurement. *Monitoring Point* is a specialization of Geographic Point used to represent named geographic points. *Monitoring Point Name* is a specialization of UFO-A Abstract Individual used to describe the location of the monitoring point. *Monitoring Programs* are specializations of UFO-C

Normative Descriptions that have in their scope monitoring activities and allocate monitoring points and monitoring facilities to perform them. *Monitoring Facilities* are stations or platforms composed of monitoring devices that directly and repeatedly measure environmental properties. Monitoring facilities are artificial geographic features. *Monitoring Devices* are specializations of Device. *Monitoring Point Principal*, *Monitoring Program Principal*, and *Monitoring Facility Principal* are used to represent the agents responsible for monitoring points, monitoring programs, and monitoring facilities, respectively. They are specializations of UFO-C Agents.

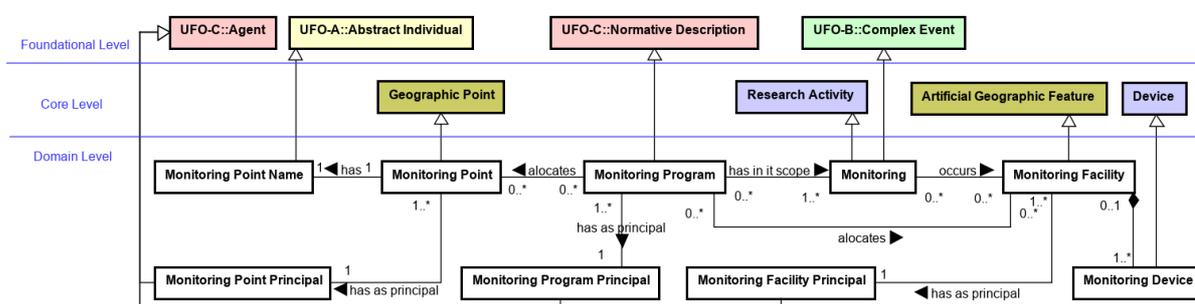


Figure 34 - The Environmental Monitoring Ontology.

Table 41 presents the Environmental Monitoring Facilities UML Model of INSPIRE elements whose reuse resulted in each *Environmental Monitoring Ontology* concept.

Table 41 - Correspondences between Environmental Monitoring Ontology concepts and Environmental Monitoring Facilities UML Model of INSPIRE elements

Environmental Monitoring Ontology concept	Environmental Monitoring Facilities UML Model element
Monitoring	EnvironmentalMonitoringActivity
Monitoring Facility	EnvironmentalMonitoringFacility
Monitoring Device	EnvironmentalMonitoringFacility
Monitoring Facility Principal	responsibleParty
Monitoring Program	EnvironmentalMonitoringProgramme
Monitoring Program Principal	responsibleParty
Monitoring Point	representativePoint
Monitoring Point Name (new concept)	-
Monitoring Point Principal	responsibleParty

6.4.2 The Water Quality Ontology

The *Water Quality Ontology* comprises concepts about water quality entities, properties and normative. Figure 35 presents this ontology. A *Water Quality Entity*, a specialization of UFO-

A Substantial, can be a *Hydrographic Feature*, a *Quantity of Water*, a *Quantity of Sediment*, a *Water Treatment Plant*, etc. Hydrographic Feature is a specialization of Natural Geographic Feature and represents rivers, lakes, hydrographic basins, seas, wells, etc. Hydrographic Feature is divided into *Hydrographic Basin* that can be simple or complex, *Surface Water*, *Sea* and *Ground Water*. Surface water is water on the surface of continents such as in a river and lake. Groundwater is the water present beneath Earth's surface in soil pore spaces and in the fractures of rock formations. *River* and *Lake* are specializations of surface water. *Well* is a specialization of groundwater. The concepts related to hydrographic feature are based on the Hydrography UML Model of INSPIRE.

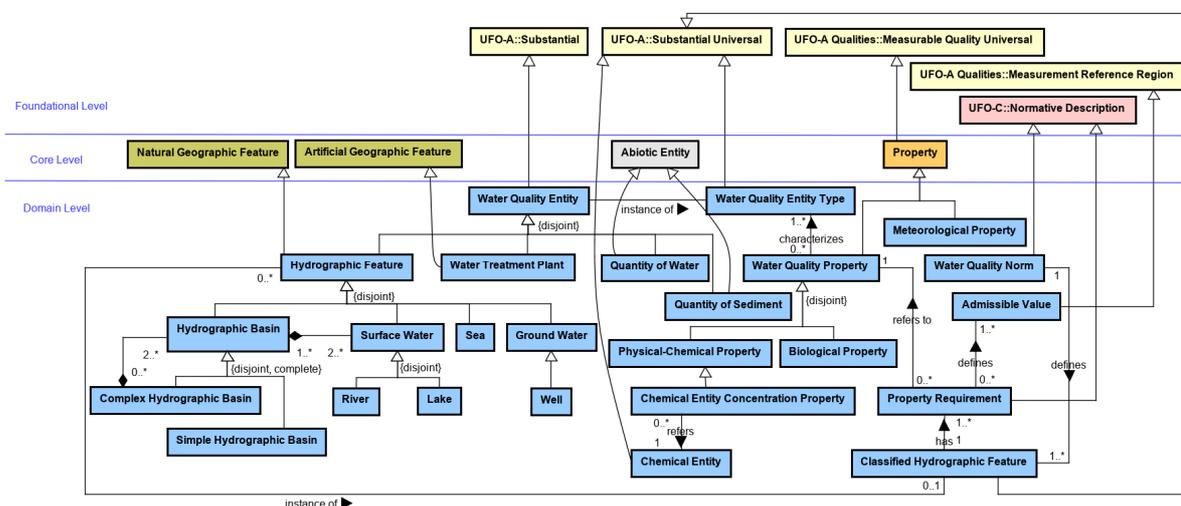


Figure 35 - The Water Quality Ontology.

Water Quality Property, a specialization of UFO-A Quality Universal, deals with properties that are used to characterize water quality entities, encompassing both *Physical-Chemical* (e.g., temperature, dichloroethene concentration) and *Biological Properties* (e.g., concentration of coliforms, algae). *Chemical Entity Concentration Property* is a specialization of *Physical-Chemical Property*. It refers to a *Chemical Entity*, a specialization of UFO-A Substantial Universal, such as calcium carbonate. The concept of chemical entity is based on the chemical entity of ChEBI Molecular Structure Ontology. However, here it is modeled as universal rather than individual as we need to take as instance the types of existing chemical entities. *Meteorological Property* is a specialization of Property that represents meteorological aspects (for instance, the amount of rain over a given period).

Water Quality Norm is a specialization of UFO-C Normative Requirement Description (e.g., 357/2005 CONAMA Resolution). It classifies a hydrographic feature according to a set of

Property Requirements (that are also specializations of UFO-C Normative Description). A *Classified Hydrographic Feature* is a specialization of UFO-A Substantial Universal that represents the classification assigned to a hydrographic feature. For example, 357/2005 CONAMA Resolution defines the class “Freshwater - Class 1” for freshwater that may be intended for: human consumption; protection of aquatic communities; primary contact recreation such as swimming, water skiing and diving; etc. A *Property Requirement* defines a Water Quality Property and *Admissible Values* for this property. For a hydrographic feature instantiates a classified hydrographic feature, it must comply with the admissible values for the water quality properties required by that classification. For instance, the class “Freshwater - Class 1” of 357/2005 CONAMA Resolution sets the maximum value of 10 µg/L for the property chlorophyll a.

Table 42 presents the knowledge resources elements whose adapted reuse resulted in each *Water Quality Ontology* concept.

Table 42 - Correspondences between Water Quality Ontology concepts and knowledge resources reused elements

Water Quality concept	Knowledge Resource reused	Knowledge Resource reused element
Water Quality Entity (new concept)	-	-
Hydrographic Feature	Hydrography UML Model of INSPIRE	HydroObject
Hydrographic Basin	Hydrography UML Model of INSPIRE	DrainageBasin
Simple Hydrographic Basin (new concept)	-	-
Complex Hydrographic Basin (new concept)	-	-
Surface Water	Hydrography UML Model of INSPIRE	SurfaceWater
River (new concept)	-	-
Lake (new concept)	-	-
Sea	Hydrography UML Model of INSPIRE	SeaArea
Ground Water (new concept)	-	-
Well (new concept)	-	-
Water Treatment Plant (new concept)	-	-
Quantity of Water	EnvO Material terms	liquid water
Quantity of Sediment	EnvO Material terms	sediment
Water Quality Entity Type (new concept)	-	-
Water Quality Property (new concept)	-	-
Physical-Chemical Property (new concept)	-	-

Chemical Entity Concentration Property (new concept)	-	-
Chemical Entity	ChEBI Molecular Structure Ontology	chemical entity
Biological Property (new concept)	-	-
Meteorological Property (new concept)	-	-
Water Quality Norm (new concept)	-	-
Classified Hydrographic Feature (new concept)	-	-
Property Requirement (new concept)	-	-
Admissible Value (new concept)	-	-

6.5 Evaluation of the Ontology Network

In this section, the proposed ontology network is evaluated. This is done through ontology verification and validation activities from NeOn [10]. For the ontology network verification, we check if the elements of the ontology network (concepts, relations, and properties) answer each of the integration questions defined in Table 14. For the ontology network validation, we show how the elements of the data sources to be integrated are represented by the ontology network elements and present some instances of them.

6.5.1 Verification of the Ontology Network

Table 43 lists the elements of the ontology network (concepts, relations, and properties) needed to answer each of the integration questions defined in Table 14. As can be seen, all integration questions faced by domain experts are answered by the ontology network elements.

Table 43 - Checking the ontology network elements that answer the integration questions

Integration Question	Ontology Network Concepts, Relations, and Properties
<p>IQ01: Which monitoring points have appropriate bathing conditions according to the analysis of thermotolerant coliforms?</p> <p>According to 274/2000 CONAMA Resolution [66], places with thermotolerant coliforms > 2500/100mL are improper for bathing.</p>	<p>Biological Property is subtype of Water Quality Property Water Quality Property is subtype of Property</p> <p>Property is supertype of Measured Property Measurement /measures Measured Property Measurement results Measured Value</p> <p>Measurement locates Geographic Point or Measured Entity is subtype of Researchable Entity Material Entity is subtype of Researchable Entity Sample is subtype of Material Entity</p>

	<p>Sampling results Sample</p> <p>Sampling locates Geographic Point</p> <p>Geographic Point is supertype of Monitoring Point</p>																				
<p>IQ02: What is the relation between upstream sewage treatment and concentration of thermotolerant coliforms?</p>	<p>Biological Property <i>is subtype of</i> Water Quality Property</p> <p>Water Quality Property <i>is subtype of</i> Property</p> <p>Property <i>is supertype of</i> Measured Property</p> <p>Measurement <i>/measures</i> Measured Property</p> <p>Measurement <i>results</i> Measured Value</p> <p>Measurement <i>locates</i> Geographic Point</p> <p>or</p> <p>Measured Entity <i>is subtype of</i> Researchable Entity</p> <p>Material Entity <i>is subtype of</i> Researchable Entity</p> <p>Sample <i>is subtype of</i> Material Entity</p> <p>Sampling <i>results</i> Sample</p> <p>Sampling <i>locates</i> Geographic Point</p>																				
<p>IQ03: Which parameters present concentrations above the thresholds established in the applicable legislation for freshwater (357/2005 CONAMA Resolution class 1)?</p>	<p>Water Quality Norm <i>defines</i> Classified Hydrographic Feature</p> <p>Classified Hydrographic Feature <i>has</i> Property Requirement</p> <p>Property Requirement refers to Water Quality Property</p> <p>Property Requirement <i>defines</i> Admissible Value</p> <p>Water Quality Property <i>is subtype of</i> Property</p> <p>Property <i>is supertype of</i> Measured Property</p> <p>Measurement <i>/measures</i> Measured Property</p> <p>Measurement <i>results</i> Measured Value</p>																				
<p>IQ04: What is the Water Quality Index (WQI) at each monitored point?</p> <p>According to [14], WQI can be calculated by:</p> <p style="text-align: center;">Table 44 - Weights assigned to parameters for WQI calculation extracted from [14]</p> <table border="1" data-bbox="228 1435 691 1890"> <thead> <tr> <th>Parameter - qi</th> <th>Weight - wi</th> </tr> </thead> <tbody> <tr> <td>Dissolved Oxygen (%DOSat)</td> <td>0.17</td> </tr> <tr> <td>Thermotolerant Coliforms* (NMP/100ml)</td> <td>0.15</td> </tr> <tr> <td>pH</td> <td>0.12</td> </tr> <tr> <td>Biochemical Oxygen Demand (mg/L)</td> <td>0.10</td> </tr> <tr> <td>Nitrate (mg/L NO₃⁻)</td> <td>0.10</td> </tr> <tr> <td>Total Phosphate (mg/L PO₄⁻²)</td> <td>0.10</td> </tr> <tr> <td>Temperature Range (°C)</td> <td>0.10</td> </tr> <tr> <td>Turbidity (UNT)</td> <td>0.8</td> </tr> <tr> <td>Total Solids (mg/L)</td> <td>0.8</td> </tr> </tbody> </table> <p>*Replaced by <i>E. coli</i> from 2013.</p> $WQI = \prod_{i=1}^9 q_i^{w_i}$ <p>Where:</p>	Parameter - qi	Weight - wi	Dissolved Oxygen (%DOSat)	0.17	Thermotolerant Coliforms* (NMP/100ml)	0.15	pH	0.12	Biochemical Oxygen Demand (mg/L)	0.10	Nitrate (mg/L NO ₃ ⁻)	0.10	Total Phosphate (mg/L PO ₄ ⁻²)	0.10	Temperature Range (°C)	0.10	Turbidity (UNT)	0.8	Total Solids (mg/L)	0.8	<p>Physical-Chemical Property <i>is subtype of</i> Water Quality Property</p> <p>Biological Property <i>is subtype of</i> Water Quality Property</p> <p>Water Quality Property <i>is subtype of</i> Property</p> <p>Property <i>is supertype of</i> Measured Property</p> <p>Measurement <i>/measures</i> Measured Property</p> <p>Measurement <i>results</i> Measured Value</p> <p>Measurement <i>locates</i> Geographic Point</p> <p>or</p> <p>Measured Entity <i>is subtype of</i> Researchable Entity</p> <p>Material Entity <i>is subtype of</i> Researchable Entity</p> <p>Sample <i>is subtype of</i> Material Entity</p> <p>Sampling <i>results</i> Sample</p> <p>Sampling <i>locates</i> Geographic Point</p> <p>Geographic Point <i>is supertype of</i> Monitoring Point</p>
Parameter - qi	Weight - wi																				
Dissolved Oxygen (%DOSat)	0.17																				
Thermotolerant Coliforms* (NMP/100ml)	0.15																				
pH	0.12																				
Biochemical Oxygen Demand (mg/L)	0.10																				
Nitrate (mg/L NO ₃ ⁻)	0.10																				
Total Phosphate (mg/L PO ₄ ⁻²)	0.10																				
Temperature Range (°C)	0.10																				
Turbidity (UNT)	0.8																				
Total Solids (mg/L)	0.8																				

<ul style="list-style-type: none"> • WQI = Water Quality Index, ranging from 1 to 100 • q_i = quality of parameter i • w_i = weight assigned to parameter i 	
<p>IQ05: What is the relation between meteorological and seasonal conditions and water quality?</p>	<p>Meteorological Property <i>is subtype of</i> Property Physical-Chemical Property <i>is subtype of</i> Water Quality Property Biological Property <i>is subtype of</i> Water Quality Property Water Quality Property <i>is subtype of</i> Property</p> <p>Property <i>is supertype of</i> Measured Property Measurement <i>/measures</i> Measured Property Measurement <i>results</i> Measured Value</p> <p>Measurement <i>locates</i> Geographic Point or Measured Entity <i>is subtype of</i> Researchable Entity Material Entity <i>is subtype of</i> Researchable Entity Sample <i>is subtype of</i> Material Entity Sampling <i>results</i> Sample Sampling <i>locates</i> Geographic Point</p>
<p>IQ06: What is the relation between river flow and water quality?</p>	<p>Physical-Chemical Property <i>is subtype of</i> Water Quality Property Biological Property <i>is subtype of</i> Water Quality Property Water Quality Property <i>is subtype of</i> Property</p> <p>Property <i>is supertype of</i> Measured Property Measurement <i>/measures</i> Measured Property Measurement <i>results</i> Measured Value</p> <p>Measurement <i>locates</i> Geographic Point or Measured Entity <i>is subtype of</i> Researchable Entity Material Entity <i>is subtype of</i> Researchable Entity Sample <i>is subtype of</i> Material Entity Sampling <i>results</i> Sample Sampling <i>locates</i> Geographic Point</p>
<p>IQ07: What is the BOD (Biochemical Oxygen Demand) / COD (Chemical Oxygen Demand) ratio at the monitoring points?</p>	<p>Physical-Chemical Property <i>is subtype of</i> Water Quality Property Property Water Quality Property <i>is subtype of</i> Property</p> <p>Property <i>is supertype of</i> Measured Property Measurement <i>/measures</i> Measured Property Measurement <i>results</i> Measured Value</p> <p>Measurement <i>locates</i> Geographic Point or Measured Entity <i>is subtype of</i> Researchable Entity Material Entity <i>is subtype of</i> Researchable Entity Sample <i>is subtype of</i> Material Entity Sampling <i>results</i> Sample Sampling <i>locates</i> Geographic Point</p>

<p>IQ08: Was there metal contamination at the collection sites prior to the incident?</p>	<p>Chemical Entity Concentration Property <i>is subtype of</i> Physical-Chemical Property Physical-Chemical Property <i>is subtype of</i> Water Quality Property Water Quality Property <i>is subtype of</i> Property</p> <p>Property <i>is supertype of</i> Measured Property Measurement <i>/measures</i> Measured Property Measurement <i>results</i> Measured Value</p> <p>Measurement <i>locates</i> Geographic Point Measurement <i>begin</i> Time Point Measurement <i>end</i> Time Point or Measured Entity <i>is subtype of</i> Researchable Entity Material Entity <i>is subtype of</i> Researchable Entity Sample <i>is subtype of</i> Material Entity Sampling <i>results</i> Sample Sampling <i>locates</i> Geographic Point Sampling <i>begin</i> Time Point Sampling <i>end</i> Time Point</p> <p>Geographic Point <i>is supertype of</i> Monitoring Point</p>
<p>IQ09: Is there contamination by metals in samples collected after the incident? How much of this contamination is past tense?</p>	<p>Chemical Entity Concentration Property <i>is subtype of</i> Physical-Chemical Property Physical-Chemical Property <i>is subtype of</i> Water Quality Property Water Quality Property <i>is subtype of</i> Property</p> <p>Property <i>is supertype of</i> Measured Property Measurement <i>/measures</i> Measured Property Measurement <i>results</i> Measured Value</p> <p>Measured Entity <i>is subtype of</i> Researchable Entity Material Entity <i>is subtype of</i> Researchable Entity Sample <i>is subtype of</i> Material Entity Sampling <i>results</i> Sample Sampling <i>locates</i> Geographic Point Sampling <i>begin</i> Time Point Sampling <i>end</i> Time Point</p> <p>Geographic Point <i>is supertype of</i> Monitoring Point</p>
<p>IQ10: Do the levels of metals found exceed the values proposed by the legislation?</p>	<p>Water Quality Norm <i>defines</i> Classified Hydrographic Feature Classified Hydrographic Feature <i>has</i> Property Requirement Property Requirement <i>refers to</i> Water Quality Property Property Requirement <i>defines</i> Admissible Value Chemical Entity Concentration Property <i>is subtype of</i> Physical-Chemical Property Physical-Chemical Property <i>is subtype of</i> Water Quality Property</p>

	<p>Water Quality Property <i>is subtype of</i> Property</p> <p>Property <i>is supertype of</i> Measured Property Measurement <i>/measures</i> Measured Property Measurement <i>results</i> Measured Value</p>
IQ11: Do sediment metal levels exceed thresholds adopted by environmental agencies?	<p>Water Quality Norm <i>defines</i> Classified Hydrographic Feature Classified Hydrographic Feature <i>has</i> Property Requirement Property Requirement refers to Water Quality Property Property Requirement <i>defines</i> Admissible Value Chemical Entity Concentration Property <i>is subtype of</i> Physical-Chemical Property Physical-Chemical Property <i>is subtype of</i> Water Quality Property Water Quality Property <i>is subtype of</i> Property</p> <p>Property <i>is supertype of</i> Measured Property Measurement <i>/measures</i> Measured Property Measurement <i>results</i> Measured Value</p> <p>Measured Entity <i>is subtype of</i> Researchable Entity Material Entity <i>is subtype of</i> Researchable Entity Abiotic Entity <i>is subtype of</i> Material Entity Quantity of Sediment <i>is subtype of</i> Abiotic Entity</p>
IQ12: Do the collected water samples present toxicity? IQ13: What types of toxicity of the water samples?	<p>Water Quality Property <i>is subtype of</i> Property</p> <p>Property <i>is supertype of</i> Measured Property Measurement <i>/measures</i> Measured Property Measurement <i>results</i> Measured Value</p> <p>Measured Entity <i>is subtype of</i> Researchable Entity Material Entity <i>is subtype of</i> Researchable Entity Sample <i>is subtype of</i> Material Entity Abiotic Entity <i>is subtype of</i> Material Entity Quantity of Water <i>is subtype of</i> Abiotic Entity</p>
IQ14: Is toxicity related to contamination levels?	<p>Chemical Entity Concentration Property <i>is subtype of</i> Physical-Chemical Property Physical-Chemical Property <i>is subtype of</i> Water Quality Property Water Quality Property <i>is subtype of</i> Property</p> <p>Property <i>is supertype of</i> Measured Property Measurement <i>/measures</i> Measured Property Measurement <i>results</i> Measured Value</p>

6.5.2 Validation of the Ontology Network

Table 45 maps the ontology network concepts that represent the elements of the data sources to be integrated and shows some instances of the ontology network concepts. They were extracted from Table 15 (Renova Foundation) and Table 16 (IBAMA-IEMA, and IGAM).

This shows how existing water quality data can be integrated using the proposed ontology network.

Table 45 - Checking the ontology network concepts that represent the elements of the data sources to be integrated

Data Source	Data Source Element	Ontology/Concept	Instance
Renova Foundation	Data Provider	Measurement/Research Activity Principal	Renova Foundation
	Period	Measurement/Time Point	28-Jan-2019 to 03-Feb-2019
	Telemetric Stations	Environmental Monitoring/Monitoring Facility	RCA 02
	Water Course	Water Quality/River	Carmo River
	Cyanobacteria (µg/L)	Water Quality/Biological Property	Cyanobacteria
		Measurement/Measure Unit	µg/L
	Electric Conductivity (µS/cm)	Measurement/Measured Value	0.4
		Water Quality/Physical-Chemical Property	Electric Conductivity
		Measurement/Measure Unit	µS/cm
	Dissolved Oxygen (mg/L)	Measurement/Measured Value	73.7
		Water Quality/Physical-Chemical Property	Dissolved Oxygen
	pH	Measurement/Measure Unit	mg/L
		Measurement/Measured Value	8.6
	Rain of the period (mm)	Water Quality/Physical-Chemical Property	pH
Measurement/Measured Value		8.4	
Water Quality/Meteorological Property		Rain of the period	
IBAMA-IEMA	Measurement/Measure Unit	mm	
	Measurement/Measured Value	0.0	
	Data Provider	Sampling or Measurement/Research Activity Principal	IBAMA-IEMA
	Site	Spatial Location/Administrative Unit	MG
	Sample Point Short Name	Environmental Monitoring/Monitoring Point	AFL-06
	Sample Point Long Name	Environmental Monitoring/Monitoring Point Name	Piranga MG - Upstream
	Sample Point Category	Material Entity/Abiotic Entity	Lotic fresh water
	Lat	Spatial Location/Coordinate	-20.383574
	Long	Spatial Location/Coordinate	-42.902283
	X	Spatial Location/Coordinate	718948
	Y	Spatial Location/Coordinate	7744747
	Z	Spatial Location/Coordinate	
	Projection	Spatial Location/Coordinate System	UTM23S
	Datum	Spatial Location/Datum	SIRGAS2000
	Date	Sampling/Time Point	10-Mar-2016 11:00
	Sample Ref	Sampling/Sample	62277-2016
	Lab Ref	Sampling/Sample	62277-2016
	Data Source	Measurement/Research Activity Agent	Merieux
	Sample Type	Measurement/Research Activity Procedure	Superficial
	Alkalinity of bicarbonates (mgCaCO3/L)	Water Quality/ Chemical Entity Concentration Property	Alkalinity of bicarbonates
Water Quality/Chemical Entity		CaCO3	
Measurement/Measure Unit		mgCaCO3/L	
Measurement/Measured Value	30.6		
IGAM	Data Provider	Measurement/Research Activity Principal	IGAM
	Hydrographic Basin	Water Quality/Hydrographic Basin	Doce River
	Sub Basin	Water Quality/Hydrographic Basin	Piranga River
	UPGRH	Spatial Location/Region	DO1 - Piranga River
	County	Spatial Location/Administrative Unit	PIRANGA (MG)
	Water Course	Water Quality/River	Piranga River
	Description	Environmental Monitoring/Monitoring Point Name	Piranga River in the city of Piranga
	Framing Class of Water Course	Water Quality/Classified Hydrographic Feature	Class 2
	Station	Environmental Monitoring/Monitoring Facility	RD001
	Altitude	Spatial Location/Coordinate	610
	Latitude (Decimal Degrees)	Spatial Location/Coordinate	-20.69
	Spatial Location/Coordinate Unit	Decimal Degrees	

Latitude (Degrees Minutes Seconds)	Spatial Location/Coordinate	-20° 41' 18.661"
	Spatial Location/Coordinate Unit	Degrees Minutes Seconds
Longitude (Decimal Degrees)	Spatial Location/Coordinate	-43.3
	Spatial Location/Coordinate Unit	Decimal Degrees
Longitude (Degrees Minutes Seconds)	Spatial Location/Coordinate	-43° 18' 8.42"
	Spatial Location/Coordinate Unit	Degrees Minutes Seconds
Year	Sampling/Time Point	2017
Sampling Date	Sampling/Time Point	02-Jul-2017
Sampling Time	Sampling/Time Point	09:15:00
Alkalinity of bicarbonates	Water Quality/Chemical Entity Concentration Property	Alkalinity of bicarbonates
	Water Quality/Chemical Entity	CaCO3
	Measurement/Measured Value	18.8

6.6 Related Work

In this section, we discuss existing models for integrating water quality data (section 6.6.1) and models used to represent scientific research activities in general (section 6.6.2), as this is a central aspect of the proposed ontology network.

6.6.1 Models for the Integration of Water Quality Data

The application of CLeAR to the water quality domain has revealed some works focused on the construction of models (e.g., ontologies) for the integration of water quality data. These models were not selected for reuse because the INSPIRE conceptual model [53] was rated better in CLeAR cycle III. In Chapter 5, we discuss the INSPIRE conceptual model. Below we briefly present these other models.

The water quality vocabulary proposed by [43] and [44] includes an observable property ontology inspired by O&M but aligned with existing ontologies. By formalizing this ontology, and clearly labelling the separate concerns, water quality observations from different sources may be more easily merged and also transformed to O&M for cross-domain applications. However, this ontology focuses on measurements, properties, units of measure, material entities, and sensors, but does not deal with other domain aspects such as spatial and temporal location, geographic entities, meteorological aspects, agents, normative, and so on.

The SSN-based ontology for water quality management, called InAWaterSense, presented by [45] and [46] supports water quality classification based on different regulation authorities. This ontology addresses measurements, properties, units of measurement, spatial and temporal location, geographic entities, material entities, sensors, and normative. It does not represent other types of research activities like sampling and monitoring, meteorological aspects and agents. Only the computational representation of this ontology is provided. Data represented from them can be accessed via a web portal [67].

The ontology-based system proposed by [47] has the intent of providing semantic interoperability for environmental monitoring data. This system is based on the Modular Environmental Monitoring Ontology (MEMOn) to represent the knowledge about the environmental domain. Unlike previous ontologies, MEMOn is grounded on the foundational ontology Basic Formal Ontology (BFO) [48]. In addition, MEMOn reuse other ontologies (e.g., SSN, EnvO). It does not address all types of research activities (e.g., preparation, monitoring), research activities methods and normative elements.

The Observation Data Model (ODM) presented by [68] provides a format for the storage and retrieval of environmental observations made at a point in a relational database designed to facilitate integrated analysis of large data sets collected by multiple investigators. This model is used to enable the publication of research datasets consisting of observations made at a point [69].

Two other related works, identified outside the systematic search, are web portals for the publication of water quality data. The Water Quality Portal (WQP) [70] is a cooperative service sponsored by the United States Geological Survey (USGS), the Environmental Protection Agency (EPA), and the National Water Quality Monitoring Council (NWQMC) for water quality monitoring data. It serves data collected by over 400 state, federal, tribal, and local agencies. In turn, the Water Quality Archive [71] provides data on water quality measurements carried out by the Environment Agency of UK Government. The first provides a water quality exchange data model. The second provides documentation on the structure of data in this archive, and the meanings of the terms used. We were not able to identify whether they are based on some ontology.

6.6.2 Models related to Scientific Research Activities

There are some models [68][72][73][74][75] related to scientific research activities based on the Observations and Measurements conceptual model from ISO 19156 (O&M) [51]. As presented earlier, O&M defines an observation as an activity, the result of which is an estimate of the value of a property of the feature of interest, obtained using a specified procedure. Specializations of the observation have been classified by the result-type. For example, a measurement is an observation whose result is a scaled quantity, and a truth observation is an observation whose result is a Boolean value. As well as in O&M, the ontologies proposed by [72][73] do not represent the sampling activity; they represent only

the sampling features. A sampling feature is used to support the observation process and may or may not have a persistent physical expression. Physical samples are modeled as the sampling feature specimen. Just like O&M, [72] implements sample preparation using an association class with specimen. As sampling is not modeled as an activity, sampling properties need to be assigned to other entities. Specimen has properties related to sampling time, sampling location, etc. Observation has phenomenon time and result time to differentiate the moment of the sampling from the time of the ex-situ measurement of a sample, respectively. Thus, events and objects concepts are mixed. This shows the importance of developing core and domain ontologies based on a foundational ontology, characteristic not presented by these models.

The Semantic Sensor Network (SSN) ontology [74] describes sensors and their observations, the involved procedures, the studied features of interest, the samples used to do so, and the observed properties, as well as actuators. In SSN, the sampling activity is modeled. Sampling is used to represent both sampling and preparation activities. Location is not addressed. It is suggested that other models must be used to deal with location. Agents and devices involved in observations are treated by the same sensor entity. The Extensible Observation Ontology (OBOE) [75] is a formal ontology for capturing the semantics of scientific observation and measurement. OBOE does not handle other research activities. The Observation Data Model (ODM) [68] represents observation results, sample properties, monitoring locations, but does not model the research activities themselves, which is key to capturing provenance information.

6.7 Concluding Remarks

In this chapter, we have designed a network of reference ontologies for the integration of water quality data. Unlike related work, the proposed ontology network covers all domain aspects identified from the application of CLeAR to the water quality domain. We show that this makes it possible to answer the integration questions faced by domain experts. We also show that this enables mapping the correspondence between the elements of the data sources to be integrated and the concepts of the ontology network. As a consequence, data from different data sources can be integrated from the shared conceptualization addressed by the ontology network. Thus, we can say that the ontology network serves the purpose for which it was built.

Regarding the organization of the ontology network in a layered architecture, we realize that the adoption of UFO as a foundational ontology supports the correct classification of the different concepts and relations, leveraging key notions that are domain independent. Activities are modeling as events, actors as agents, devices as objects, their participations in events revealed, and so on. By not adhering to a foundational ontology, some misconceptions arise, e.g., with event properties assigned to objects as verified in related work.

Moreover, in this work, the adoption of a foundational ontology has enabled the reengineering and integration of previous knowledge resources from incompatible formats. For instance, concepts provided by EnvO such as material entity, available in OWL and OBO formats, can be used in conjunction with concepts provided by the O&M conceptual model such as feature of interest, available in UML. By reusing existing structured resources we avoid unnecessary proliferation of new knowledge resources.

Still in relation to the ontology network architecture, we can point out that the reuse of the core level ontologies facilitates the domain level ontologies development process. This can be verified from the *Environmental Monitoring Ontology* in which concepts of the *Spatial Location Ontology* and the *Scientific Research Activity Ontology* were specialized to represent concepts related to environmental monitoring. Thus, the concepts provided by the core level ontologies can be specialized to expand the ontology network by including new domain level ontologies (e.g., an air quality ontology).

It is noteworthy that the core level ontologies modeled in this work provide concepts that can be reused for modeling domain level ontologies from other areas of knowledge, since concepts related to material entities, spatial location and research activities are not specific of the environmental domain. For example, to represent the health care domain, it is necessary to speak about urine and blood samples, measurements of properties related to these material entities, location of origin of pathologies, etc. That is, it is necessary to address the subjects covered by these ontologies.

Finally, we can say that the explicit modeling of research activity reveals that provenance information, usually present in the metadata domain, are actually properties of events, including the participation of agents and non-agentive objects in those events. In the case of scientific research, the modeling of these concepts is fundamental to support the integrated data reuse. Otherwise, there is a risk that such data will be misused. For example,

data produced by incompatible methods can be compared, leading to inconsistent analysis; incorrect providers can be assigned to data since original data can be reprocessed by different agents; and so on. In the next chapter, we discuss final considerations.

7 Final Considerations

This final chapter presents the main contributions of this work and the future research directions. Section 7.1 presents a summary and the main contributions of this work. Section 7.2 shows the applicability of CLeAR and the proposed ontology network in other scenarios. Section 7.3 presents the limitations and difficulties faced in carrying out the work. Finally, section 7.4 discusses future work.

7.1 Summary of the Work

Enabling data-centric environmental science, management and decision-making requires proper support for data semantics. In this work, we have addressed this challenge for environmental data in the Doce River Basin, building a network of reference ontologies for the integration of water quality data. The ontology network spans several domain aspects such as research activities; methods and devices used to perform these activities; actors involved; spatial location; material entities analyzed; water quality properties checked (physical, chemical and biological properties); etc.

As we intended to reuse existing knowledge resources in the construction of the ontology network to avoid unnecessary proliferation of new ontologies, we sought a reuse-oriented ontology engineering methodology. We chose the NeOn methodology. None of the reuse-oriented ontology engineering methodologies consulted (including NeOn) addresses the search and selection of reusable knowledge resources systematically. As a consequence, we have decided to develop CLeAR to deal with these activities in a systematic way.

CLeAR addresses this gap in ontology engineering methodologies by applying some practices of the Systematic Literature Review to find existing knowledge resources about a scientific research domain. In addition, CLeAR evaluates the knowledge resources found according to domain coverage and some objective quality attributes. Finally, CLeAR is aligned to the needs of ontology building for the purpose of scientific research data integration, since the scope of the ontology is derived from integration questions faced by domain experts and data to be integrated.

As a result, CLeAR provides a set of evaluated and classified structured resources on a scientific research domain. In this work, the application of CLeAR to the water quality

domain has resulted in a knowledge base with 75 structured resources. Besides CLeAR and the proposed ontology network, this is a relevant contribution since the set of knowledge resources can be revisited. This should help offset the effort to apply CLeAR.

Regarding the development of CLeAR, the analysis of related work shows that there are other initiatives trying to solve the problem of “integrating environmental data” based on the reuse of existing knowledge resources. Each of these initiatives builds their ontology from previous structured resources but without adopting a systematic approach. This further motivates us to face the problem of reusing existing knowledge resources for the development of ontologies using systematic methods. High quality shared ontology models can enhance the information production and its accuracy, especially in cases in which data sources are produced in a heterogeneous way.

In relation to the knowledge resources found and selected for reuse, we verified that none of them addresses all aspects covered by the water quality domain, since the spectrum of aspects is very wide. Because of this, it was necessary to reuse different knowledge resources in an integrated manner. However, as the knowledge resources are produced in incompatible formats, they cannot be integrated into their original form. As we adopt UFO to ground the ontology network, we have used UFO to analyze and adapt the elements of knowledge resources so that they could be integrated into the ontology network.

We realize that the adoption of a foundational ontology is a key feature of the proposed ontology network because it supports the correct classification of different concepts and relations. Activities are modeled as events, actors as agents, devices as objects, their participations in events revealed, and so on. As discussed in related work, by not adhering to a foundational ontology, some misconceptions arise, e.g., with event properties assigned to objects.

A central fragment of the proposed ontology network is the *Scientific Research Activity Ontology*. The explicit modeling of research activities reveals that provenance information, usually present in the metadata domain, are actually properties of events, including the participation of agents and non-agentive objects in those events. In the case of scientific research, the modeling of these concepts is fundamental to support integrated data reuse.

Finally, we have demonstrated how the ontology network can provide integrated semantics to water quality data. To do so, we show the concepts of the ontology network that address each of the integration questions identified by the domain experts and the correspondence between the elements of the data sources to be integrated and the ontology network concepts.

7.2 Applicability of the Work in other Scenarios

The CLeAR approach is aimed at finding knowledge resources to be reused in the development of ontologies for the purpose of integrating scientific research data. Although developed in the context of the Doce River Project to address a need related to the environmental domain, the approach provides guidelines that are free from a specific domain, as explained in Chapter 3. This way CLeAR can be applied to the different domains in which the integration of scientific research data is required, such as health care research.

One of the fragments of the resulting ontology network demonstrates this potential, namely the extension of the Core Ontology on Measurement (COM) with concepts related to the sampling activity [76]. The extension of this core ontology was possible because the systematic search returned some relevant publications and structured resources related to the aspects of measurement and sampling. These aspects compose the environmental domain, but are also relevant to the representation of other domains as presented in [76].

Regarding the proposed ontology network, although we have considered integrating water quality data from the Doce River Basin to build it, this ontology network applies to water quality research in general, and hence has the potential to benefit integration efforts in many other scenarios. Besides that, due to the architecture adopted for the construction of the ontology network, new specializations can be made from the core level ontologies so that other environmental subdomains can be addressed (e.g., air quality).

Finally, the core ontologies can be reused in the development of other domain level ontologies that involve material entities, spatial location and research activities. In particular, the *Scientific Research Activity Ontology*, which deals with the different types of research activities performed in empirical research, can be reused to model any domain where empirical research is performed.

7.3 Limitations and Difficulties

The main limitation regarding CLeAR refers to the availability of the structured resources identified from its application to a specific domain. During the course of this work, some of the knowledge resources resulted from the application of CLeAR to the water quality domain were discontinued, others were turned into commercial products. This makes reusing these knowledge resources unfeasible.

Among the difficulties encountered in performing this work, we can mention the bureaucracy faced to obtain data to be integrated. In many cases, such data are not available online. Thus, it was in many cases necessary to contact each provider for access. Another difficulty identified was the lack of documentation or examples of use of some reusable structured resources. Documentation and examples are essential for the activities of verifying domain coverage, understanding the knowledge resources, and aligning them with a foundational ontology. If they are not available, the effort to carry out these activities, which is not small, increases considerably.

7.4 Future Work

The designed ontology network forms the basis of mechanisms for finding, publishing and querying heterogeneous environmental data. Based on the ontology network, a semantic data repository can be built and evolved into a public portal for water quality data for the Doce River Basin. Besides that, data extractors capable of translating the tabular data from several data sources can be built. The repository in this portal can provide researchers and the general public with access to data that would otherwise be poorly accessible and hard to integrate.

Also regarding the integration and availability of heterogeneous water quality data, we note the need to define standard structures for data sources based on the proposed ontology network and to offer these structures to data producers. This could decrease data sources heterogeneity.

In relation to the ontology network coverage, some improvements can be made to broaden its scope. With respect to the core level ontologies, other types of research activities can be modeled (direct observations, complex assays, etc.). In addition, other aspects of scientific research as well as other types of research activities may be incorporated. Examples are scientific research purpose, scientific research planning, etc. Regarding the ontology

network as whole, other environmental subdomains (e.g., air quality, observation of the taxon of an animal) can be represented from new specializations of core level ontologies.

Still in relation to the ontology network, another possible improvement is the identification of the origin of the concepts coming from other knowledge resources. In this work, we show the origin of concepts only through traceability tables.

Finally, as future work related to CLeAR, we can consider evaluating the degree of coverage of domain aspects (not covered, covered, largely covered, and fully covered) rather than just whether or not they are covered by knowledge resources. We can also look for new quality attributes to be evaluated for the classification and selection of existing knowledge resources. Besides that, we can study the automation of some steps of CLeAR to reduce the effort required to apply it. As an example, we can try to automate the application of the inclusion and exclusion criteria of publications and structured resources.

8 References

- [1] GIBERT, K., HORSBURGH, J. S., ATHANASIADIS, I. N., and HOLMES G., "**Environmental Data Science**," *Environmental Modelling and Software*, 2018, vol. 106, p. 4-12.
- [2] UHLIR, P. F., and SCHRÖDER, P., "**Open Data for Global Science**," *Data Science Journal*, 2007, vol. 6, p. 36-53.
- [3] LENZERINI, M., "**Data Integration: A Theoretical Perspective**," in *Proceedings of the 21st ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS 2002)*, 2002, p. 233-246.
- [4] RAJPATHAK, D., and CHOUGULE, R., "**A generic ontology development framework for data integration and decision support in a distributed environment**," *International Journal of Computer Integrated Manufacturing*, 2011, vol. 24, p. 154-170.
- [5] CRUZ, I. F., and XIAO, H., "**The Role of Ontologies in Data Integration**," *Journal of Engineering Intelligent Systems*, 2005, vol. 13, p. 245-252.
- [6] GRUBER, T. R., "**The Role of Common Ontology in Achieving Sharable, Reusable Knowledge Bases**," in *Principles of knowledge representation and reasoning: Proceedings of the Second International Conference*, 1991.
- [7] ASHBURNER, M. *et al.*, "**Gene ontology: Tool for the unification of biology**," *Nature Genetics*, 2000, vol. 25, p. 25-29.
- [8] USCHOLD, M., HEALY M., WILLIAMSON K., CLARK, P., and WOODS S., "**Ontology reuse and application**," in *Formal Ontology in Information Systems*, IOS Press, 1998.
- [9] BONTAS, E. P., MOCHOL, M., and TOLKSDORF, R., "**Case Studies on Ontology Reuse**," in *Proceedings of the IKNOW05 International Conference on Knowledge Management*, 2005..
- [10] SUÁREZ-FIGUEROA, M. C., GÓMEZ-PÉREZ, A., MOTTA E., and GANGEMI, A., "**Ontology engineering in a networked world**," 2012.
- [11] FALBO, R. A., "**SABiO: Systematic approach for building ontologies**," in *1st Joint Workshop Onto.Com/ODISE on Ontologies in Conceptual Modeling and Information Systems Engineering*, 2014.
- [12] GUARINO, N., "**Formal Ontology and Information Systems**," *Formal ontology in information systems: Proceedings of the first international conference*, IOS Press, 1998.
- [13] FERNANDES, G. W. *et al.*, "**Deep into the mud: ecological and socio-economic impacts of the dam breach in Mariana, Brazil**," *Natureza e Conservação*, 2016, vol. 14, p. 35-45.
- [14] ANA, "**Agência Nacional de Águas**," 2019. [Online]. Available: <https://www.ana.gov.br/>. [Accessed: 07-Jun-2019].

- [15] CPRM, "**Serviço Geológico do Brasil**," 2019. [Online]. Available: <http://www.cprm.gov.br/>. [Accessed: 07-Jun-2019].
- [16] IBAMA, "**Instituto Brasileiro do Meio Ambiente e dos Recursos Naturais Renováveis**," 2019. [Online]. Available: <https://www.ibama.gov.br/>. [Accessed: 07-Jun-2019].
- [17] IGAM, "**Instituto Mineira de Gestão das Águas**," 2019. [Online]. Available: <http://www.igam.mg.gov.br/>. [Accessed: 07-Jun-2019].
- [18] IEMA, "**Instituto Estadual de Meio Ambiente e Recursos Hídricos**," 2019. [Online]. Available: <https://iema.es.gov.br/>. [Accessed: 07-Jun-2019].
- [19] RENOVA F., "**Fundação Renova**," 2019. [Online]. Available: <https://www.fundacaorenova.org/>. [Accessed: 07-Jun-2019].
- [20] HEY, T., and TREFETHEN, A. E., "**Cyberinfrastructure for e-Science**," *Science*, 2005, vol. 308, p. 817-821.
- [21] WILKINSON, M. D. *et al.*, "**Comment: The FAIR Guiding Principles for scientific data management and stewardship**," *Scientific Data*, 2016, vol. 3.
- [22] DYBA, T., KITCHENHAM, B. A., and JORGENSEN, M., "**Evidence-based software engineering for practitioners**," *IEEE Software*, 2005, vol. 22, p. 58-65.
- [23] KITCHENHAM, B., and CHARTERS, S., "**Guidelines for performing Systematic Literature reviews in Software Engineering Version 2.3**," *Engineering*, 2007.
- [24] GUIZZARDI, G., "**Ontological foundations for structural conceptual models**," *PhD Thesis*, University of Twente, The Netherlands, 2005.
- [25] GUIZZARDI, G., FALBO, R. A., and GUIZZARDI, R. S. S., "**Grounding Software Domain Ontologies in the Unified Foundational Ontology (UFO): The case of the ODE Software Process Ontology**," in *Proceedings of the 11th Iberoamerican Conference on Software Engineering (CibSE 2008)*, 2008, p. 127-140.
- [26] ALBUQUERQUE, A., and GUIZZARDI, G., "**An ontological foundation for conceptual modeling datatypes based on semantic reference spaces**," in *Proceedings of the International Conference on Research Challenges in Information Science*, 2013, p. 1-12.
- [27] SCHERP, A., SAATHOFF, C., FRANZ, T., and STAAB, S., "**Designing core ontologies**," *Applied Ontology*, 2011, vol. 6, p. 177-221.
- [28] RUY, F. B., FALBO, R. D. A., BARCELLOS, M. P., COSTA, S. D., and GUIZZARDI, G., "**SEON: A software engineering ontology network**," in *Proceedings of the 20th International Conference on Knowledge Engineering and Knowledge Management (EKAW'16)*, 2016.
- [29] GUIZZARDI, G., "**On Ontology, ontologies, Conceptualizations, Modeling Languages, and (Meta)Models**," in *Proceedings of the 2007 conference on Databases and Information Systems IV: Selected Papers from the Seventh International Baltic Conference (DB&IS'2006)*, 2007, p. 18-39.

- [30] STUDER, R., BENJAMINS, V. R., and FENSEL, D., "**Knowledge Engineering: Principles and methods**," *Data & Knowledge Engineering*, 1998, vol. 25, p. 161-197.
- [31] D'AQUIN, M., SCHLICHT, A., STUCKENSCHMIDT, H., and SABOU, M., "**Ontology Modularization for Knowledge Selection: Experiments and Evaluations**," in *Database and Expert Systems Applications: 18th International Conference*, 2007, p. 874-883.
- [32] GÓMEZ-PÉREZ, A., FERNÁNDEZ-LÓPEZ, M., and CORCHO, O., "**Ontological Engineering with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web**," Springer Verlag, 2010.
- [33] SIMPERL, E., MOCHOL, M., BURGER, T., and POPOV, I. O., "**Achieving maturity: The state of practice in ontology engineering in 2009**," *International Journal of Computer Science and Applications*, 2009, vol. 7, p. 45-65.
- [34] SOARES A., "**Towards ontology-driven information systems: Guidelines to the creation of new methodologies to build ontologies**," *PhD Dissertation*, The Pennsylvania State University, Ann Arbor, 2009.
- [35] GRUNINGER, M., and FOX, M. S., "**Methodology for the Design and Evaluation of Ontologies**," in *International Joint Conference on Artificial Intelligence (IJCAI95), Workshop on Basic Ontological Issues in Knowledge Sharing*, 1995, p. 1-10.
- [36] KITCHENHAM, B. A., BUDGEN, D., and PEARL BRERETON, O. , "**Using mapping studies as the basis for further research - A participant-observer case study**," *Information and Software Technology*, 2011, vol. 53, p. 638-651.
- [37] PETERSEN, K., FELDT, R., MUJTABA, S., and MATTSSON, M., "**Systematic Mapping Studies in Software Engineering**," in *Proceedings of the 12th International Conference on Evaluation and Assessment in Software Engineering (EASE'08)*, 2008, p. 68-77.
- [38] WOHLIN, C., "**Guidelines for snowballing in systematic literature studies and a replication in software engineering**," in *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering (EASE '14)*, 2014.
- [39] CONAMA, "**Resolução n° 357, de 17 de março de 2005**," 2005.
- [40] RICE, E. W., BAIRD, R. B., and EATON A. D., "**Standard methods for the examination of water and waste water**," American Public Health Association, American Water Works Association, Water Environment Federation, 2017.
- [41] CAMPOS, P. M. C. *et al.*, "**Building an ontology network to support environmental quality research: First steps**," in *Proceedings of the XI Seminar on Ontology Research in Brazil and II Doctoral and Masters Consortium on Ontologies*, 2018, p. 227-232.
- [42] CAMPOS, P. M. C., REGINATO, C. C., and ALMEIDA, J. P. A., "**Application of the CLeAR Approach to the Water Quality Domain**," Mendeley Data, 2019. [Online]. Available: <https://data.mendeley.com/datasets/kx3wrhpwnb/draft?a=e3e6d14c-5403-4cdf-ae83-e26b85026228>. [Accessed: 04-Oct-2019].
- [43] COX, S. J. D., SIMONS, B. A., and YU, J., "**A Harmonized Vocabulary For Water**

- Quality,"** in *HIC2014 - 11th International Conference on Hydroinformatics*, 2014.
- [44] SIMONS, B. A., YU, J., and COX, S. J. D., "**Defining a water quality vocabulary using QUDT and ChEBI,**" in *Proceedings of the 20th International Congress on Modelling and Simulation (MODSIM)*, 2013, pp. 2548-2554.
- [45] AHMEDI, L., JAJAGA, E., and AHMEDI, F., "**An ontology framework for water quality management,**" in *Proceedings of the 6th International Conference on Semantic Sensor Networks (SSN'13)*, 2013, vol. 1063, p. 35-50.
- [46] JAJAGA, E., AHMEDI, L., and AHMEDI, F. , "**An expert system for water quality monitoring based on ontology,**" in *Communications in Computer and Information Science*, 2015.
- [47] MASMOUDI, M. *et al.*, "**An ontology-based monitoring System for multi-source environmental observations,**" in *Procedia Computer Science*, 2018, vol. 126, p. 1865-1874.
- [48] ARP, R., SMITH, B., and SPEAR, A. D., "**Building Ontologies with Basic Formal Ontology,**" MIT Press, 2016.
- [49] RUY, F. B., "**Software Engineering Standards Harmonization: An Ontology-Based Approach,**" *PhD Thesis*, Federal University of Espírito Santo, 2017.
- [50] BARCELLOS, M. P., FALBO, R. A., and FRAUCHES, V. G. V., "**Towards a measurement ontology pattern language,**" in *in Proceedings of the 1st Joint Workshop ONTO.COM/ODISE on Ontologies in Conceptual Modeling and Information Systems Engineering*, 2014.
- [51] ISO, ISO 19156:2011, "**Geographic information - Observations and measurements,**" 2011.
- [52] HODGSON, R., KELLER, P. J., HODGES, J., and SPIVAK, J., "**QUDT - Quantities, Units, Dimensions and Data Types Ontologies,**" *W3C*, 2014.
- [53] INSPIRE, "**Infrastructure for Spatial Information in Europe,**" 2019. [Online]. Available: <https://inspire.ec.europa.eu/>. [Accessed: 07-Jun-2019].
- [54] INSPIRE Thematic Working Group Hydrography, "**D2.8.I.8 Data Specification on Hydrography – Technical Guidelines,**" 2014.
- [55] INSPIRE Thematic Working Group Administrative Units, "**D2.8.I.4 Data Specification on Administrative Units – Technical Guidelines,**" 2014.
- [56] INSPIRE Thematic Working Group Environmental Monitoring Facilities, "**D2.8.II/III.7 Data Specification on Environmental Monitoring Facilities – Technical Guidelines,**" 2013.
- [57] TOM H., and ROSWELL C., "**Standards Guide ISO/TC 211 GEOGRAPHIC INFORMATION/GEOMATICS,**" 2009.
- [58] ISO, ISO 19111:2007, "**Geographic Information: Spatial referencing by coordinates,**" 2006.
- [59] BUTTIGIEG, P. L., MORRINSON, N., SMITH, B., MUNGALL, C. J., and LEWIS, S.

- E., **"The environment ontology: Contextualising biological and biomedical entities,"** *Journal of Biomedical Semantics*, 2013.
- [60] BUTTIGIEG, P. L., PAFILIS, E., LEWIS, S. E., SCHILDHAUER, M. P., WALLS, R. L., and MUNGALL, C. J., **"The environment ontology in 2016: Bridging domains with increased scope, semantic density, and interoperability,"** *Journal of Biomedical Semantics*, 2016.
- [61] EnvO, **"Environment Ontology,"** 2019. [Online]. Available: <http://environmentontology.org/>. [Accessed: 07-Oct-2019].
- [62] HASTINGS, J. *et al.*, **"ChEBI in 2016: Improved services and an expanding collection of metabolites,"** *Nucleic Acids Research*, 2016.
- [63] BBSRC, **"Chemical Entities of Biological Interest (ChEBI),"** 2019. [Online]. Available: <https://www.ebi.ac.uk/chebi/>. [Accessed: 29-Jul-2019].
- [64] W3C, **"W3C Basic Geo (WGS84 lat/long) Vocabulary,"** *SWIG Basic Geo (WGS84 lat/long) Vocabulary*, 2006.
- [65] PERRY, M., and HERRING, J., **"OGC GeoSPARQL-A geographic query language for RDF data,"** 2012.
- [66] CONAMA, **"Resolução nº 274, de 29 de novembro de 2000,"** 2000.
- [67] AHMEDI, L., SEJDIU, B., BYTYÇI, E., and AHMEDI, F., **"An integrated web portal for water quality monitoring through wireless sensor networks,"** *Int. J. Web Portals*, 2015.
- [68] HORSBURGH, J. S., TARBOTON, D. G., MAIDMENT, D. R., and ZASLAVSKY, I., **"A relational model for environmental and water resources data,"** *Water Resources Research*, 2008.
- [69] HORSBURGH, J. S. *et al.*, **"An integrated system for publishing environmental observations data,"** *Environmental Modelling & Software*, 2009, vol. 24, p. 879-888
- [70] READ, E. K. *et al.*, **"Water quality data for national-scale aquatic research: The Water Quality Portal,"** *Water Resources Research*, 2017.
- [71] Environment Agency, **"Water Quality Archive,"** 2019.
- [72] COX, S. J. D. , **"An explicit OWL representation of ISO/OGC observations and measurements,"** in *Proceedings of the 6th International Conference on Semantic Sensor Networks (SSN'13)*, 2013, vol. 1063, p. 1-18.
- [73] COX, S. J. D., **"Ontology for observations and sampling features, with alignments to existing models,"** *Semantic Web*, 2017, vol. 8, p. 453-470.
- [74] HALLER, A. *et al.*, **"The Modular SSN Ontology: A Joint W3C and OGC Standard Specifying the Semantics of Sensors, Observations, Sampling, and Actuation,"** *Semantic Web*, 2018, vol. 10, p. 9-32.
- [75] MADIN, J., BOWERS, S., SCHILDHAUER, M., KRIVOV, S., PENNINGTON, D. P., and VILLA, F., **"An ontology for describing and synthesizing ecological observation data,"** *Ecological Informatics*, 2007, vol. 2, p. 279-296.

- [76] SANTOS, L., BARCELLOS, M. P., FALBO, R. A., REGINATO, C. C., and CAMPOS, P. M. C. "**Measurement Task Ontology**," in *Proceedings of the XII Seminar on Ontology Research in Brazil*, 2019.