Finding Reusable Structured Resources for the Integration of Environmental Research Data

Patricia M. C. Campos[1], Cássio C. Reginato[1], João Paulo A. Almeida[1], Monalessa P. Barcellos[1], Ricardo de Almeida Falbo[1], Vítor E. Silva Souza[1], Giancarlo Guizzardi[1,2]

[1]Ontology & Conceptual Modeling Research Group (NEMO),
Federal University of Espírito Santo (UFES),
Av. Fernando Ferrari, 514, Goiabeiras, 29075-910, Vitória, ES, Brazil
[2]Conceptual and Cognitive Modeling Research Group (CORE),
Free University of Bozen-Bolzano,
Bolzano, Italy

patmarcal@gmail.com, {cassio.reginato, jpalmeida,
monalessa, falbo, vitorsouza, gguizzardi}@inf.ufes.br

Abstract. *Successful data integration requires careful examination of data semantics, a task that has often been approached with the use of ontologies. 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. This paper addresses this challenge by proposing a systematic approach (dubbed CLeAR) for the identification and selection of reusable artifacts for building ontologies with the purpose of research data integration. CLeAR follows some principles of the systematic literature reviews, supporting the search for structured resources in the scientific literature. We apply CLeAR to the environmental domain. A total of 543 publications were surveyed. The results obtained provide a set of 75 structured resources for the environmental domain, evaluated according domain coverage and some quality attributes (e.g., proper documentation, community acceptance).*

Keywords: *Data integration; environmental research data; knowledge resources; reuse; systematic search; ontology.*

1 Introduction

Scientific research is often a data-centric endeavor, involving the systematic collection, interpretation and evaluation of scientific data [1]. In several domains, scientific research comprises: (i) the interaction between many actors (such as academic institutions, government agencies, private companies and independent research groups), (ii) carrying out research activities (such as observation and measurement), and (iii) the use of various nomenclatures and classification schemes (types of materials collected, types of properties observed, etc.). In these settings, scientific data is produced from a variety of sources, in different contexts and for a variety of purposes. As a consequence, such data is produced in heterogeneous forms.

Given the high costs involved in producing scientific data (e.g., for environmental data science [2]), it is no surprise that significant gains can be obtained

from data sharing, reuse and integration [3]. Data integration demands strategies to deal with data heterogeneity whether in terms of syntax, schema or semantics [4]: *syntactic heterogeneity* occurs mainly 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 the same data can be used. The semantic aspect, which is the focus of this paper, has been frequently approached with the use of ontologies [5].

As presented in [6][7], 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, 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 [8] 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 (ambiguities and inconsistencies) among existing ontologies. Thus, reuse has becoming a common concern in the ontology engineering area [9][10].

Some ontology engineering methodologies describe specific activities to deal with reuse [11][12][13]. Despite that, some challenges still need to be tackled to promote reuse. The NeOn Methodology [11], for example, proposes eight scenarios for building ontologies from the reuse of previous knowledge resources. However, such methodology 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 realized the need to propose an approach to do so in a systematic way.

The approach is dubbed CLeAR (Conducting Literature Search for Artifact Reuse) and is based on some practices of the Systematic Literature Review (SLR) [14][15]. 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 consist 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. As CLeAR addresses specific ontology engineering activities, it is designed to be used as a complement to existing ontology engineering methodologies.

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 reused whenever necessary. This justifies the effort employed (the

proposed work is not automated) in performing the systematic search for a domain for the first time.

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.

The paper is further structured as follows. Section 2 presents some background knowledge that supports our investigation on the development of an approach to search and select reusable knowledge resources for the integration of scientific research data. Section 3 describes the CLeAR approach. Section 4 discusses the results of the application of CLeAR to the environmental research domain. Finally, section 5 presents the final considerations.

2 Background

In this section, we review ontology engineering methodologies, gaps of existing methodologies related to reuse, and the Systematic Literature Review (SLR) [14][15], required for the development of this work.

2.1 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" [16]. 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 (or ontology developers) have expertise in fields such as knowledge representation and development tools, and users apply the ontology for a particular purpose [17].

In [16], the authors differentiate three types of activities within an ontology engineering process: management, development and support activities. 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 [17].

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 [18]. Examples of methodologies that address building ontologies from scratch can be found in [18]. Examples of methodologies that

describe specific activities for addressing reuse are the NeOn Methodology [11], the Systematic Approach for Building Ontologies (SABiO) [12], and the Methodology of Integration-oriented Ontology Development (MIOD) [13].

2.1.1 Reuse-Related Gaps

Reuse is pointed out as a promising approach to ontology engineering, since it enables speeding up the ontology development process, saving time and money [19], and avoids the unnecessary proliferation of new ontologies. However, there is a lack of concern with search and selection of reusable knowledge resources by the reuse-oriented ontology engineering methodologies. This is shown in [20], in which a systematic mapping was performed to provide the current panorama of ontology integration approaches. The results reveal some problems, among them, a lack of concern with search and selection of the ontologies to be integrated.

Among the reuse-oriented methodologies, some focus on the identification and the integration of existing knowledge resources (NeOn [11], SABiO [12] and MIOD [13]). In general, they propose steps for the specification of ontology requirements, for the identification of the knowledge resources to be reused, for the integration of the knowledge resources (reengineering, alignment, merging, etc.) and for the evaluation of the resulting ontology. Ontology requirements are specified mainly in the form of competency questions (CQs), i.e., questions writing in natural language that the ontology should be able to answer [21]. In turn, the terms whose definition could be reusable from existing knowledge resources are those appearing in the ontology requirements specification. Ontology developers can locate knowledge resources in ontology libraries, domain-related sites, resources within organizations, and general-purpose search engines.

Some reuse-oriented methodologies focus only on the identification of relevant knowledge resources ([22] and [23]). In [22], the authors propose an ontology pattern classification scheme to allow the reuse of existing ontology knowledge for multiagent systems development. In [23], a systematic literature review is carried out to obtain security ontologies. These ontologies are compared according to the evaluation framework proposed in [24], making it possible to identify the key requirements that an integrated security ontology should have. Other reuse-oriented methodologies focus only on the integration of two or more knowledge resources (for example, [25][26]). They assume that knowledge resources are identified in a previous step.

Besides that, there are some ontology engineering methodologies focused on data integration, but which do not address the reuse of knowledge resources. This is the case of the Methodology for Development on Data Integration (OntoDI) [27]. It proposes specific steps to identify data sources to be integrated and to correct semantic inconsistences between them.

Despite proposing activities to identify and integrate existing knowledge resources, NeOn [11], SABiO [12] and MIOD [13] do not show how to perform the search and record the search results. Regarding knowledge resources selection, MIOD suggests some evaluation criteria (for example, quality of documentation and language used to implement the resource) but does not show how to assess these criteria. NeOn applies a subjective evaluation criterion that is the consensus about the knowledge and terminology used by the resource. SABiO does not describe how knowledge resources

are to be selected. In relation to the methodologies [22] and [23], they search for specific types of knowledge resources (ontology patterns or ontologies) or in specific domains (security). In their turn, the methodologies proposed in [25] and [26] do not address the search for reusable knowledge resources. Finally, OntoDI [27] does not address a step related to reuse.

2.2 Systematic Literature Review

As we have discussed in the previous section, there is explicit support for reuse in ontology engineering methodologies. However, they provide only generic guidelines for reusable knowledge resources search and selection activities. This justifies a more systematic approach to perform them. We draw inspiration for such approach from the practices of the Systematic Literature Review (SLR) [14][15].

The Systematic Literature Review (SLR) [14][15] 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. 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 methodology used and the results obtained. Although an SLR requires considerable effort to be implemented when compared to ad hoc literature reviews, SLRs are auditable, more trustworthy and rigorous.

An SLR has three phases: planning, conducting and reporting the review [15]. 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 further to make 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 [15].

As a way to enhance the quality of the search, Snowballing can be performed [28]. Snowballing refers to using the reference list of a study or the citations to the study to identify additional studies, and therefore increase coverage of relevant literature. 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 section 3.

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.

The proposed approach adopts some practices of the Systematic Literature Review (SLR) [14][15]. 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, available representation and community acceptance). As a result, the set of retrieved structured resources make up a knowledge base on the domain to be 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. Consequently, it is designed to be used as a complement to existing ontology engineering methodologies. For example, when used together with NeOn [11], CLeAR activities correspond to (and replace) NeOn's specification of ontology requirements, search for reusable knowledge resources, assessment of candidate knowledge resources, and selection of knowledge resources. The overview of CLeAR activities is presented in the sequel.

3.1 Overview of CLeAR Activities

CleAR is structured in three cycles as shown in Figure 1. 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.

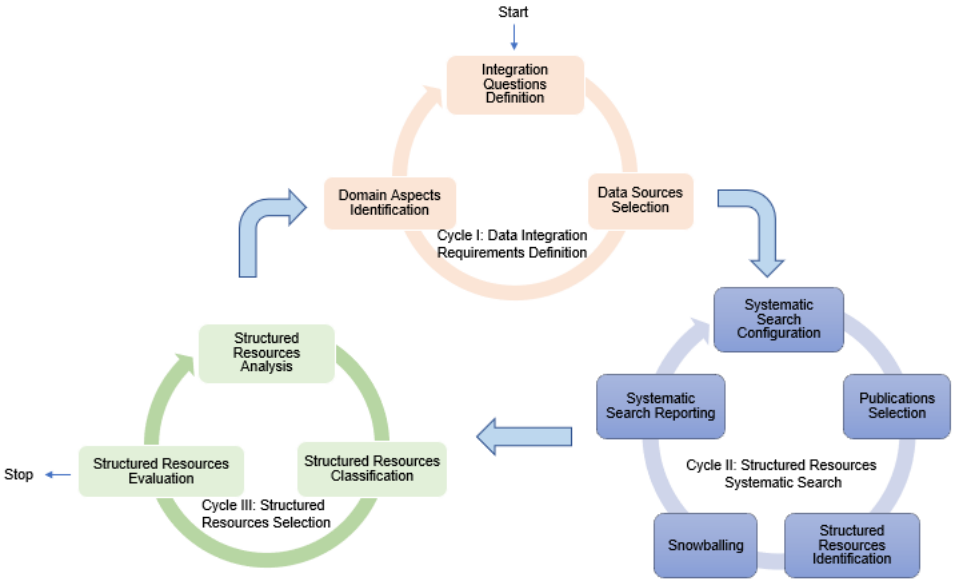

Figure 1 - CLeAR activities.

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 [4]. 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. 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 (norms, national and international standards, guides, etc.) combined with the knowledge of domain experts, allows the ontology engineers to identify the domain aspects. Domain aspects are subjects of the domain that can be treated in a modular way. They must be enough to represent the universe of discourse. They can be related to activities, actors and roles description, characterization of researched entities, and so on. 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. 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 set of 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, 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

From IQs, it is possible to define the final set of data sources, selecting those that provide appropriate data to answer IQs. The selected data sources will be integrated with the support of the ontology to be developed from the reuse of the discovered 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, the domain aspects are identified. For this, one can use general questions to characterize a scientific research that needs to consume integrated data. 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.

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 data sources 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

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 [28] 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. The domain related terms are obtained from discussions with domain experts, glossaries prepared by them, domain standards and domain aspects (when they are 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 [15]. Checking search engines results against lists of already known primary studies, called here control papers, can be useful for selection of the search engines [15].

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 [15]. 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. Secondly, one must define data to be registered about the studied publications and the form (for example, a spreadsheet) to be used to record them. Regarding publications data, it is necessary to register: the year, the title, the authors and the source.

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 studied 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 used to build the resource (such as Ontology Web Language - OWL, Extensible Markup Language – XML and Unified Modeling Language - UML), 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.

Snowballing Procedure Definition. In this step, the process to be followed for the Snowballing application is defined. 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	
Search Goals Specification	
Inputs	The motivations for the systematic search
Outputs	The systematic search goals
Keywords Selection	
Inputs	The systematic search goals
Outputs	List of keywords related to structured resources, and list of keywords related to domain
Search String Improvement	
Inputs	List of keywords related to structured resources, and list of keywords related to domain
Outputs	Search string
Search Engines Selection	
Inputs	List of control papers
Outputs	Search engines selected
Inclusion and Exclusion Criteria Definition	
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
Publications Selection Procedure Definition	
Inputs	The systematic search goals
Outputs	Process to be followed for the publications selection, form to record publications data
Structured Resources Identification Procedure Definition	
Inputs	The systematic search goals
Outputs	Process to be followed for the structured resources identification, form to record structured resources data
Snowballing Procedure Definition	
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 such as the language used to build the resources, the number of publications that mention the resources and the number of resources that reuse them. This is useful in evaluating the quality attributes of the structured resources performed in cycle III as it will be presented below. 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 and community acceptance). 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 existing 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.

3.4.1 Structured Resources Analysis

Domain Coverage Analysis. Domain coverage is analyzed based on the domain aspects. This can be verified by checking whether or not a domain aspect is covered by structured resources or indicating a degree of coverage. 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 row 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																		
	How			Where			When			What			Who			Why		
Structured Resource Name	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							✓	✓	✓	✓	✓	✓						
...	✓	✓	✓															

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) [11]. 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 (through the creation of a conceptual model that describes the solution to a problem). The second one is because it provides a machine-readable implementation version of the structured resource. We have used the language used to build 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 row 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.

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	✓	✓	✓	✓			4
SR05		✓	✓	✓	✓	✓	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. 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

4 Applying CLeAR to the Water Quality Domain

In this section, 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 an ontology for the integration of water quality data. The work was carried out by two domain experts and two ontology engineers over a period of 2 months. Cycle I and cycle II activities took approximately 2 weeks each and cycle III activities took approximately 1 month. The most time-consuming step is the Structured Resources Analysis in cycle III, as it is necessary to study each of the identified structured resources to verify the domain coverage and quality attributes. The domain experts are researchers in the areas of Geochemistry and Aquatic Biodiversity. The ontology engineers already had knowledge about the water quality domain before applying the approach, which reduced the time required to study publications and structured resources. It is worth mentioning that the time required for applying CLeAR depends directly on the number of publications and structured resources to be analyzed, as well as the size of the structured resources and the quality of documentation available on them. In turn, the number of publications and structured resources to be analyzed is driven by the requirements specified in cycle I, that is, IQs (more generic or specific), data sources to be integrated and domain aspects.

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, who are knowledgeable of data semantics and who face themselves integration questions in their research activities.

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 could be 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) [29], the Geological Survey of Brazil (CPRM) [30] and the Brazilian Institute for the Environment and Renewable Natural Resources (IBAMA) [31]. The state-level governmental organizations selected are: the Water Management Institute of Minas Gerais (IGAM) [32] and the Institute of Environment and Water Resources of Espírito Santo (IEMA) [33]. The non-governmental organization selected is Renova Foundation [34], that is the entity responsible for the mobilization to repair damages caused by the rupture of the Fundão dam, in Mariana (MG).

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 [34]. 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* (µg/L, µS/cm, mg/L and mm); *sensors used*

(telemetric stations); *reference to norms* (357/2005 CONAMA Resolution [35] 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 (µg/L)**			**Electric Conductivity (µS/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-Fev-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* (mgCACO3/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., [36]) 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 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 (CP) 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 [37]), 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 (PIC) and exclusion criteria (PEC) are shown in Table 20 and the structured resources inclusion (SRIC) and exclusion criteria (SREC) 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).

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.

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.

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 [37], 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 [38] 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 [38].

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 [38].

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 [38].

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. The analysis of publications and structured resources was divided among ontology engineers, which reviewed each other’s work. Divergences in analysis were discussed and resolved in meetings.

4.2.5 Reporting the Results of the Systematic Search

As previously discussed, the systematic search returned a total of 543 publications, of which 85 (15.7%) were selected for presenting or mentioning 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 used to build the resources, the number of publications that mention these resources (not including the papers that present them) and the number of resources that reuse them. Such data is used in cycle III to evaluate the quality attributes of the structured resources. 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. Ontology Web Language (OWL) is used by 38.9% of the structured resources found while schemas written in Resource Description Framework (RDF) and Extensible Markup Language (XML) have reached 22.2%. Only 8.3% use Unified Modeling Language (UML), 6.5% use Hypertext Markup Language (HTML), in this case structured links, and 24.1% use other languages. For this analysis (see graph of Figure 2), 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.

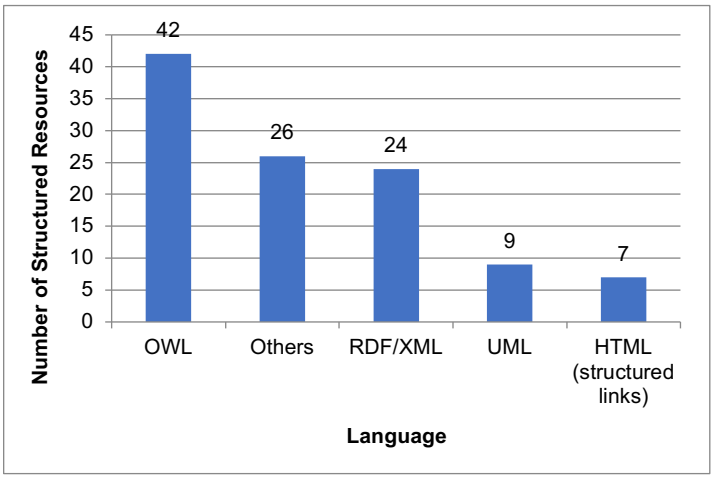

Figure 2 - Language used by the structured resources.

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 3, two structured resources, Semantic Sensor Network (SSN) Ontology [39] and Semantic Web for Earth and Environmental Terminology (SWEET) Ontologies [40], are mentioned by fourteen publications; one structured resource, the Observations and Measurements (O&M) Conceptual Model [41], is mentioned by thirteen publications; one resource, the Chemical Entities of Biological Interest (ChEBI) Ontology [42], is mentioned by ten publications; two resources, Time Ontology in OWL (OWL-Time) [43] and Quantity, Unit, Dimension and Type (QUDT) Ontologies [44], by nine publications; and one resource, Water Markup Language (WaterML) [45], 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).

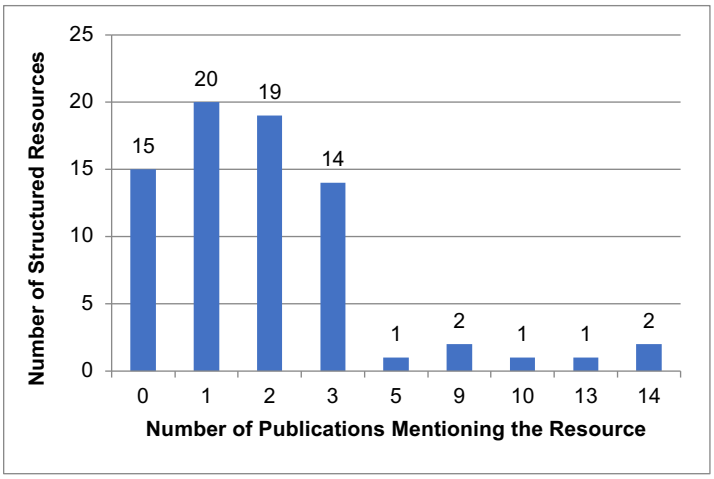

Figure 3 - Popularity of structured resources according to the number of identified publications that mention them.

The number of resources that reuse a structured resource represents how much it is used by the community. Regarding the number of resources that reuse a structured resource, the graph of Figure 4 shows that one structured resource, O&M [41], is reused by twelve resources; one structured resource, Geography Markup Language (GML) [46], is reused by eight resources; one structured resource, SSN [39], is reused by seven resources; one structured resource, the standard Geographic information/Geomatics (ISO/TC 211) [47], is reused by six resources; one structured resource, OWL-Time [43], is reused by five resources; and one structured resource, SWEET [40], 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.

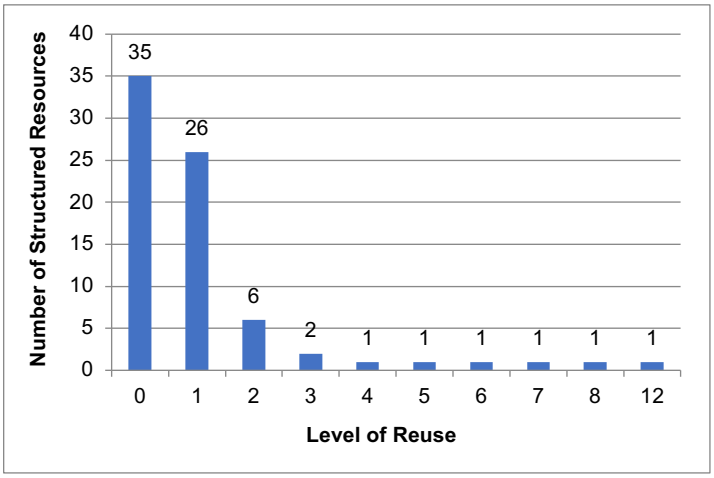

Figure 4 - Level of reuse of structured resources according to the number of structured resources that adopt them.

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 the graph of Figure 3. This is to ensure that the structured resources are recognized and reused by the community and not just by the group that have created them.

4.3 Selection of the Structured Resources on the Water Quality Domain

In this section, the application of the cycle III of CLeAR to the water quality domain is discussed.

4.3.1 Analyzing the Structured Resources

Table 22 shows the domain coverage analysis for the selected structured resources. The complete analysis was recorded in the "Structured Resources Selection" table of the dataset (which includes citations to all of the resources) [38]. In Table 22, to improve the view of the domain coverage by groups, the columns of the domain aspects that make up each group were painted with the same color. The structured resources were ordered by the total of domain aspects covered by them (from largest to smallest).

Table 22 - Structured Resources Domain Coverage Matrix

Domain Coverage																													
	How											Where			When	What									Who			Why
Structured Resource Name	Research Activities	Sampling	Preparation	Measurement	Analysis	Monitoring	Sampling Method	Preparation Method	Measurement Method	Analysis Method	Monitoring Method	Location	Geographic Coordinates	Geographic Entity	Temporal References	Material Entity	Abiotic Entity	Biotic Entity	Properties	Chemical Property	Physical Property	Biological Property	Unit of Measurement	Meteorological Aspects	Agent	Sensor	Provenance	Normative Elements
USGS Thesaurus	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓
INSPIRE	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓				✓	✓	✓	✓	✓	✓	✓
SWEET	✓	✓		✓	✓	✓						✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
GEMET	✓	✓		✓	✓	✓	✓		✓	✓	✓	✓		✓	✓	✓	✓	✓		✓	✓	✓		✓	✓	✓		✓	✓
ISO/TC 211		✓	✓	✓			✓	✓	✓			✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓		✓
UsgsHydroML		✓		✓		✓			✓			✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓
Darwin Core	✓	✓	✓	✓			✓	✓	✓			✓	✓	✓	✓	✓		✓	✓			✓	✓		✓		✓		✓
Upper Cyc	✓			✓								✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓
SUMO				✓								✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓
InAWaterSense				✓					✓			✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓			✓		✓	✓
WDTF		✓	✓	✓			✓	✓	✓			✓	✓	✓	✓	✓			✓				✓		✓		✓	✓	✓
EML	✓	✓					✓	✓	✓	✓	✓		✓		✓	✓			✓				✓		✓	✓	✓		
MEMOn		✓		✓			✓		✓			✓	✓	✓	✓	✓			✓				✓	✓	✓	✓	✓		
GeoSciML		✓	✓	✓	✓		✓	✓	✓	✓		✓			✓	✓			✓		✓				✓		✓		
EnvO		✓				✓						✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓					
GCMD				✓	✓	✓						✓	✓	✓			✓	✓		✓	✓	✓		✓	✓	✓			
WaterML		✓		✓			✓		✓			✓	✓	✓	✓	✓			✓				✓		✓	✓			
ODM		✓		✓	✓		✓		✓	✓		✓	✓		✓				✓				✓		✓	✓			
O&M		✓	✓	✓			✓	✓	✓			✓			✓	✓			✓						✓	✓			
CCO				✓								✓	✓	✓	✓	✓			✓		✓		✓		✓	✓	✓		
EIA				✓	✓									✓		✓	✓	✓	✓	✓	✓	✓		✓	✓				
EAO				✓	✓				✓	✓						✓	✓	✓	✓			✓	✓	✓	✓				
WQOP				✓					✓							✓	✓	✓	✓	✓	✓	✓	✓			✓			
OM-Heavy		✓	✓	✓			✓		✓			✓			✓	✓			✓						✓	✓			
Wavellite				✓					✓			✓	✓	✓	✓	✓			✓					✓	✓	✓			
WaWO												✓		✓	✓		✓	✓	✓	✓	✓	✓				✓			
SAM-Lite		✓		✓			✓		✓			✓			✓	✓									✓	✓	✓		
WQO				✓								✓	✓	✓					✓	✓	✓		✓			✓		✓	
WaWO+												✓	✓	✓			✓	✓	✓	✓	✓	✓						✓	
SERONTO		✓		✓			✓		✓						✓				✓				✓		✓	✓			
BCO		✓					✓					✓			✓	✓	✓	✓							✓		✓		
new SSN		✓		✓			✓		✓						✓				✓						✓	✓			
SensorML				✓					✓			✓	✓		✓				✓						✓	✓			
GML				✓					✓			✓	✓	✓	✓				✓				✓						
PEIA	✓			✓								✓		✓		✓								✓	✓			✓	
ECS												✓	✓		✓				✓				✓		✓	✓	✓		
OBOE				✓					✓										✓	✓	✓	✓	✓						
OM-Lite				✓					✓						✓				✓				✓		✓	✓			
EABS				✓					✓			✓			✓				✓						✓	✓			
Glossary BAP					✓							✓		✓		✓	✓								✓		✓		
VSTO												✓	✓		✓				✓	✓	✓					✓			
SemSOS				✓					✓			✓	✓		✓				✓										
SEGO				✓								✓			✓				✓						✓	✓			
Uberon																✓		✓	✓	✓	✓	✓							
WMO				✓					✓							✓			✓				✓	✓					
SSN				✓					✓										✓						✓	✓			
PROV-O												✓			✓	✓									✓		✓		

WSSN			✓				✓									✓					✓	✓		
QUDT																✓	✓	✓	✓	✓				
OM																✓	✓	✓	✓	✓				
QU Rec 20																✓	✓	✓		✓				
CF																✓	✓	✓			✓			
Irstea Hydro			✓				✓									✓						✓		
MMI																✓						✓	✓	
WGS84									✓	✓		✓												
FTT									✓	✓	✓													
GeoNames									✓	✓	✓													
TGN									✓	✓	✓													
USBGN									✓	✓	✓													
NGA/GNS									✓	✓	✓													
GeoSPARQL									✓	✓	✓													
QU																✓				✓				
UCUM																✓				✓				
QUDV																✓				✓				
GAZ									✓		✓													
NCBITaxon													✓		✓									
QB			✓													✓								
EngMath																		✓		✓				
MUO																		✓		✓				
OWL-Time												✓												
UO																				✓				
SWRL Temporal												✓												
MDO																					✓			
ChEBI																		✓						
DAML-Time												✓												

The structured resources positioned at the beginning of Table 22 address a greater number of domain aspects than the others. They deal with domain aspects contained in most groups, tending to be more generic, e.g., United States Geological Survey (USGS) Thesaurus [48], Infrastructure for Spatial Information in Europe (INSPIRE) [49] and SWEET [40]. The structured resources positioned at the end cover a smaller number of domain aspects, contained in one or two groups. Thus, they tend to be more specific. As examples, we can mention GeoNames [50] and GeoSPARQL [51] ("Where"); OWL-Time [43] ("When"); and QUDT [44] and ChEBI [42] ("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 [38].

Table 23 - Structured Resources Quality Attributes Matrix

Structured Resource Name	Quality Attributes						
	Proper Documentation		Available Representation		Community Acceptance		Quality Attributes Score
	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 [41] and SSN [39]) rank positively in all 6 quality attributes; two structured resources (QUDT [44] and OWL-Time [43]) 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 [38]. 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
	Darwin Core, EML	7	0.64	3	0.50	0.57
Where	GML, ISO/TC 211, WaterML, INSPIRE, UsgsHydroML	3	1.00	4	0.67	0.83
	Darwin Core, SWEET, GeoNames, TGN, GeoSPARQL, WDTF, GCMD, Upper Cyc	3	1.00	3	0.50	0.75
When	O&M	1	1.00	6	1.00	1.00
	OWL-Time	1	1.00	5	0.83	0.92

	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
What	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
	InAWaterSense, WQOP	8	0.89	2	0.33	0.61
Who	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
	MEMOn, ECS	3	1.00	2	0.33	0.67
Why	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 [49], well classified in the 6 groups; ISO/TC 211 [47] and United States Geological Survey Hydrologic Markup Language (UsgsHydroML) [52], well classified into 5 groups; and O&M [41] and SWEET [40], 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 [29], IBAMA [31] and IEMA [33], IGAM [32], CPRM [30] and Renova Foundation [34]. The first structured resource evaluated was the INSPIRE [49] 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 that 14 (23.7%) of the 59 data sources elements were represented by other structured resources reused by INSPIRE, 12 from O&M [41] and 2 from ISO/TC 2011 [47], 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 [44] and ENVironment Ontology (EnvO) [53], 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 [42], 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 [38].

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.

Table 25 - Fragment of Structured Resources Evaluation

Data Source		Structured Resource			
Data Source	**Data Source Element**	**Name**	**Concept (class)**	**Property**	**Instance**
IBAMA-IEMA	Data Provider	INSPIRE	RelatedParty	organisationName	
	Site	INSPIRE	HydroObject / AdministrativeUnits	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	AdministrativeUnits	name	
	Water Course	INSPIRE	Watercourse	geographicalName	
	Description	INSPIRE	EnvironmentalMonitoringFacility	additionalDescription	
	Framing Class of Water Course	INSPIRE	LegislationCitation		
	Station	INSPIRE	EnvironmentalMonit	name	

			oringFacility		
	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		

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.

5 Final Considerations

In this paper, we have presented CLeAR, an approach inspired by Systematic Literature Review practices to find reusable structured resources about a scientific research domain. CLeAR can be used with existing reuse-oriented ontology engineering methodologies (for example, NeOn [11] and MIOD [13]) to support the search and selection of reusable knowledge resources. CLeAR cycle I corresponds to the activity of ontology requirements specification of ontology engineering methodologies. In turn, CLeAR cycles II and III correspond to the knowledge resources identification. The structured resources selected from the application of CLeAR to a domain serve as input for the next activity of ontology engineering methodologies (the integration of the reusable knowledge resources). In addition, the set of IQs identified can be used to evaluate the resulting ontology in the same way that CQs are used by NeOn and MIOD.

The main advantage of using CLeAR is that it supports the identification of reusable knowledge resources in a systematic fashion, which is not addressed by existing ontology engineering methodologies. Another advantage is that it proposes the evaluation of reusable knowledge resources based on objective quality attributes, a feature not present in existing ontology engineering methodologies. In addition, CLeAR is aligned to the needs of ontology building for the purpose of scientific research data integration, with ontology requirements derived from IQs and data to be integrated.

A disadvantage of CLeAR is 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 reused whenever new needs about such domain arise. We argue that this result justifies the effort employed. It is important to state that the set of structured resources returned by applying CLeAR to a given domain depends on the requirements specified in cycle I. If IQs, data sources and domain aspects are changed, another set of structured resources can be obtained as

result. In any case, to build ontologies that need to address similar domain aspects, the same set of structured resources can be used, even though IQs and data sources are different.

Here, we have reported the application of CLeAR to the water quality domain. We focused on finding structured resources to be reused 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 the quality attributes proper documentation, available representation, and community acceptance, and classified based on this assessment. In the evaluation performed, 6 of the structured resources were able to jointly represent all elements of the data sources to be integrated. These structured resources were selected to be reused.

In [54], some of us report the use of CLeAR together with NeOn to build an ontology for the water quality domain using these 6 structured resources. As they differ from each other and cannot be integrated into their original format, a foundational ontology was employed in the analysis and reengineering of them. Most of the concepts represented by the designed ontology (42 out of a total of 78 concepts, i.e., 53.8%) were reused from the knowledge resources selected. This evidences the fruitfulness of CLeAR in promoting reuse.

The set of 75 structured resources resulting from the application of CLeAR to the water quality domain is available in [38] and provides an important knowledge base that can be reused. Thus, people who need to build ontologies for the water quality domain (or environmental domain) with similar domain aspects 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 a previous work (see [37]), we have conducted a non-systematic search for structured resources about the water quality domain. This search resulted in a set of 11 reusable knowledge resources. Some were already known to us, others were obtained from the analysis of various publications that we could identify. As can be seen, the number of structured resources obtained from the application of CLeAR is considerably higher than that obtained from the non-systematic search. It is important to mention that two of the knowledge resources identified by the non-systematic search (OntoBio [55] and M-OPL [56]) were not returned by CLeAR. OntoBio was published in Portuguese (therefore, it does not meet inclusion criteria), and M-OPL addresses a more general issue (measurements in general, not specifically targeted at the environmental quality domain). When comparing the approaches, we observe that the application of a systematic approach guides the search and broadens the scope of results. Moreover, we realize that CLeAR facilitates discovery of important initiatives and working groups in the field of interest.

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 is 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.

Finally, as future work, 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 examples, we can try to automate the application of the inclusion and exclusion criteria and the extraction of data from publications and structured resources. We can also try automating the domain coverage analysis and the quality attributes analysis as these steps are the most time consuming and this would greatly reduce the effort of applying the approach.

Acknowledgements

This work is partly supported by CNPq (407235/2017-5 and 312123/2017-5), CAPES (23038.028816/2016-41) and FAPES (69382549).

References

[1] ÇAPARLAR, C. Ö., and DÖNMEZ, A., **"What is Scientific Research and How Can it be Done?,"** Turkish Journal of Anaesthesiology and Reanimation, 2016, vol. 44, p. 212-218.

[2] GIBERT, K. *et al.*, **"Environmental Data Science,"** Environmental Modelling and Software, 2018, vol. 106, p. 4-12.

[3] UHLIR, P. F., and SCHRÖDER, P., **"Open Data for Global Science,"** Data Science Journal, 2007, vol. 6, p. 36-53.

[4] 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.

[5] 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.

[6] CRUZ, I. F., and XIAO, H., **"The Role of Ontologies in Data Integration,"** Journal of Engineering Intelligent Systems, 2005, vol. 13, p. 245-252.

[7] 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.

[8] ASHBURNER, M. *et al.*, **"Gene ontology: Tool for the unification of biology,"** Nature Genetics, 2000, vol. 25, p. 25-29.

[9] USCHOLD, M. *et al.,* **"Ontology reuse and application,"** in Formal Ontology in Information Systems, IOS Press, 1998.

[10] BONTAS, E. P., MOCHOL, M., and TOLKSDORF, R., **"Case Studies on Ontology Reuse,"** in Proceedings of the IKNOW05 International Conference on Knowledge Management, 2005.

[11] SUÁREZ-FIGUEROA, M. C. *et al.*, **"Ontology engineering in a networked**

world," 2012.

[12] 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.

[13] LEUNG, N. K. Y. *et al.*, **"An integration-oriented ontology development methodology to reuse existing ontologies in an ontology development process,"** in Proceedings of the 13th International Conference on Information Integration and Web-based Applications and Services, 2011, p. 174-181.

[14] DYBA, T., KITCHENHAM, B. A., and JORGENSEN, M., **"Evidence-based software engineering for practitioners,"** IEEE Software, 2005, vol. 22, p. 58-65.

[15] KITCHENHAM, B., and CHARTERS, S., **"Guidelines for performing Systematic Literature reviews in Software Engineering Version 2.3,"** Engineering, 2007.

[16] GÓMEZ-PÉREZ, A., FERNÁNDEZ-LÓPEZ, M., and CORCHO, O., **"Ontological Engeenering with examples from the areas of Knowledge Management, eCommerce and the Semantic Web,"** Springer Verlag, 2010.

[17] SIMPERL, E. *et al.*, **"Achieving maturity: The state of practice in ontology engineering in 2009,"** International Journal of Computer Science and Applications, 2009, vol. 7, p. 45-65.

[18] SOARES A., **"Towards ontology-driven information systems: Guidelines to the creation of new methodologies to build ontologies,"** Ph.D. Dissertation, The Pennsylvania State University, 2009.

[19] POVEDA-VILLALÓN, M., SUÁREZ-FIGUEROA, M. C., and GOMEZ-PEREZ, A., **"Reusing Ontology Design Patterns in a Context Ontology Network,"** in CEUR Workshop Proceedings, 2010.

[20] SALAMON, J. S., REGINATO, C. C., and BARCELLOS, M. P., **"Ontology integration approaches: A systematic mapping,"** in ONTOBRAS, 2018, pp. 161-172.

[21] GRUNINGER, M., and FOX, M. S., **"Methodology for the Design and Evaluation of Ontologies,"** Workshop on Basic Ontological Issues in Knowledge Sharing (IJCAI 1995), 1995, p. 1-10.

[22] SHIANG, C. *et al.*, **"Ontology reuse for multiagent system development through pattern classification,"** Software: Practice and Experience, 2018, vol. 48, p. 1923-1939.

[23] BLANCO, C. *et al.*, **"Basis for an integrated security ontology according to a systematic review of existing proposals,"** Computer Standards & Interfaces, 2011, vol. 33, p. 372-388.

[24] TELLO, A. L., and GÓMEZ-PÉREZ, A., **"ONTOMETRIC: A Method to Choose the Appropriate Ontology,"** J. Database Manag., 2004, vol. 15, p. 1-18.

[25] STOILOS, G. *et al.*, **"A novel approach and practical algorithms for ontology integration,"** in 17th International Semantic Web Conference, 2018, pp. 458-

476.

[26] CROW, L., and SHADBOLT, N., **"Extracting focused knowledge from the semantic web,"** International Journal of Human-Computer Studies, 2001, vol. 54, p. 155-184.

[27] YUNIANTA, A. *et. al*, **"OntoDI: The Methodology for Ontology Development on Data Integration,"** International Journal of Advance Computer Science and Applications, 2019, vol. 10, p. 160-168.

[28] 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.

[29] ANA, **"Agência Nacional de Águas,"** 2019. [Online]. Available: https://www.ana.gov.br/. [Accessed: 07-Jun-2019].

[30] CPRM, **"Serviço Geológico do Brasil,"** 2019. [Online]. Available: http://www.cprm.gov.br/. [Accessed: 07-Jun-2019].

[31] IBAMA, **"Instituto Brasileiro do Meio Ambiente e dos Recursos Naturais Renováveis,"** 2019. [Online]. Available: https://www.ibama.gov.br/. [Accessed: 07Jun-2019].

[32] IGAM, **"Instituto Mineira de Gestão das Águas,"** 2019. [Online]. Available: http://www.igam.mg.gov.br/. [Accessed: 07-Jun-2019].

[33] IEMA, **"Instituto Estadual de Meio Ambiente e Recursos Hídricos,"** 2019. [Online]. Available: https://iema.es.gov.br/. [Accessed: 07-Jun-2019].

[34] RENOVA F., **"Fundação Renova,"** 2019. [Online]. Available: https://www.fundacaorenova.org/. [Accessed: 07-Jun-2019].

[35] CONAMA, **"Resolução n° 357, de 17 de março de 2005,"** 2005.

[36] 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.

[37] CAMPOS, P. M. C. *et al.*, **"Building an ontology network to support environmental quality research: First steps,"** in CEUR Workshop Proc., 2018.

[38] [dataset] CAMPOS, P. M. C., REGINATO, C. C., and ALMEIDA, J. P. A., **"Application of the CLeAR Approach to the Water Quality Domain,"** Mendeley Data, 2020. [Online]. Available: http://dx.doi.org/10.17632/rjtgsjjgjv.1. [Accessed: 07-Jun-2020].

[39] COMPTON, M. *et al.*, **"The SSN Ontology of the W3C Semantic Sensor Network Incubator Group,"** Web Semantics: Science, Services and Agents on the World Wide Web, 2012, vol. 17.

[40] RASKIN, Robert, and PAN, M., **"Knowledge representation in the Semantic Web for Earth and Environmental Terminology (SWEET),"** Computers & Geosciences, 2005, vol. 31, p. 1119-1125.

[41] ISO, ISO 19156:2011, **"Geographic information -- Observations and**

measurements," 2011.

[42] HASTINGS, J. *et al.*, **"ChEBI in 2016: Improved services and an expanding collection of metabolites,"** Nucleic Acids Research, 2016.

[43] COX, S., and LITTLE, C., **"Time Ontology in OWL,"** 2017. [Online]. Available: https://www.w3.org/TR/owl-time/. [Accessed: 07-Jun-2019].

[44] HODGSON, R. *et al.*, **"QUDT - Quantities, Units, Dimensions and Data Types Ontologies,"** W3C, 2014.

[45] I. Zaslavsky, D. Valentine, T. Whiteaker, **"CUAHSI WaterML"**, Open Geospatial Consort. Discuss. Pap. OGC 07-041r1, 2007.

[46] ISO, ISO 19136:2007, **"Geographic information -- Geography Markup Language (GML),"** 2007.

[47] TOM H., and ROSWELL C., **"Standards Guide ISO/TC 211 GEOGRAPHIC INFORMATION/GEOMATICS,"** 2009.

[48] USGS, **"USGS Thesaurus,"** 2019. [Online]. Available: https://www2.usgs.gov/science/about/. [Accessed: 07-Jun-2019].

[49] INSPIRE, **"Infrastructure for Spatial Information in Europe,"** 2019. [Online]. Available: https://inspire.ec.europa.eu/. [Accessed: 07-Jun-2019].

[50] GeoNames, **"GeoNames,"** 2019. [Online]. Available: http://www.geonames.org/. [Accessed: 07-Jun-2019].

[51] PERRY, M., and HERRING, J., **"OGC GeoSPARQL-A geographic query language for RDF data,"** 2012.

[52] BERMUDEZ, L. E., and PIASECKI, M., **"HYDROML: Conceptual Development of a Hydrologic Markup Language,"** in 30th IAHR Congr., 2003.

[53] EnvO, **"Environment Ontology,"** 2019. [Online]. Available: http://environmentontology.org/. [Accessed: 07-Oct-2019].

[54] CAMPOS, P. M. C., **"Designing a Network of Reference Ontologies for the Integration of Water Quality Data,"** M.Sc. Thesis, Federal University of Espírito Santo, 2019. Available: https://nemo.inf.ufes.br/publications. [Accessed: 14-Feb-2020].

[55] ALBUQUERQUE, A. C. F., DOS SANTOS, J. L. C., and DE CASTRO, A. N., **"OntoBio: A biodiversity domain ontology for Amazonian biological collected objects,"** in Proc. Annual Hawaii International Conference on System Sciences, 2015, p. 3770-3779.

[56] BARCELLOS, M. P., FALBO, R. A., and FRAUCHES, V. G. V, **"Towards a measurement ontology pattern language,"** in Proc. 1st Joint Workshop ONTO.COM/ODISE, CEUR Workshop Proceedings, 2014, vol. 1301.