Building an Ontology Network to Support Environmental Quality Research: First Steps

Patricia M. C. Campos, Cássio C. Reginato, João Paulo A. Almeida, Julio Cesar Nardi, Monalessa P. Barcellos, Ricardo de Almeida Falbo, Renata S. S. Guizzardi

Ontology and Conceptual Modelling Research Group (NEMO) – Computer Science Department, Federal University of Espírito Santo (UFES), Vitória, Brazil

patmarcal@gmail.com, cassio@inf.ufes.br, jpalmeida@ieee.org, julionardi@ifes.edu.br, {monalessa, falbo, rguizzardi}@inf.ufes.br

Abstract. *Environmental Quality Research (EQR) comprises many different methods, procedures and subdomains, often requiring the integration of heterogenous data from many sources. In this paper we show the first steps in building an ontology network to support interoperability of EQR data. We present a bottom-up approach that begins by analyzing available data to capture relevant community concerns and then establish the domain coverage. We focus on identifying and reusing existent knowledge sources to cover the semantics of extant data.*

1. Introduction

Environmental Quality Research (EQR) performs observations and measurements to evaluate environmental conditions (Cox, 2017). This type of research usually has a specific purpose, like analyzing water quality at a particular site or understanding how an environmental disaster affects the quality of an ecosystem. In EQR practice, researchers make joint use of the data they produce with existing data from many sources, such as academic literature, environmental agencies, independent experts and consultants, etc. The requirements of data correlation bring up serious interoperability concerns, since each source may use different vocabularies and standards.

The FAIR Data initiative (FORCE11, 2016) presents a number of principles to address this problem. One of them establishes that metadata must meet domain-relevant community standards. This means that the produced data must be annotated with metadata that, in turn, must reference domain-relevant community standards such as reference ontologies. For several decades it has been often suggested that ontologies could be a key instrument to address semantic interoperability challenges.

In some success cases like (Ashburner et al., 2000), ontology-based models have become reference models that are used and reused in a large community, with beneficial consequences on data reuse. In other cases, however, ontologies fail in establishing de facto shareability (i.e., community acceptance) and consequently fail to support interoperability. This may have a number of reasons, including: (i) the lack of alignment between the various data sources and available ontologies, (ii) the coexistence of competing and incompatible data schemas and vocabularies, (iii) the lack of consideration of existing knowledge sources for the construction of ontologies.

Aware of such risks, we are currently investigating an ontology development bottom-up approach that can cope with the demands of data integration by leveraging existing ontologies. This effort is included in a project entitled “An eScience Infrastructure for Water Quality Management in the *Doce* River Basin” and is concerned with the integration of water quality data produced by various sources to assess the impacts of the disaster that occurred in the city of *Mariana*, in Brazil, in 2015 (the largest accident in history in volume of material dumped by mining tailings dams).

In this paper we show the first steps in the building of a network of interconnected ontologies (an ontology network, as presented in (Ruy et al., 2016)) for EQR using this approach. Section 2 presents the overview of our approach. Section 3 shows the identification and analysis of actual environmental research data to establish the representative concepts of the domain. Section 4 presents the identification of existing knowledge sources and the analysis of domain coverage by them so that they can be reused. Finally, section 5 discusses final considerations and future work.

2. The Overview of the Approach: From Data to Ontologies

Most ontology engineering methods propose an initial activity for defining the domain of enquiry, the purpose and scope of an ontology, and relevant knowledge sources (Soares, 2009). Regarding scope definition, many methods, such as NeOn (Suárez-Figueroa et al., 2012), opt for a top-down approach in which domain coverage arises from interactions between ontology developers and domain experts. As we have already discussed, for the domain of EQR, the “ontological needs” are related to the purpose of data integration. For this case, the traditional top-down approach alone does not seem to be the most appropriate alternative to define the ontology scope, since it does not address the origin of the problem, that is, the EQR data. Thus, we propose a bottom-up approach for the definition of the scope of ontologies that are aimed at data integration.

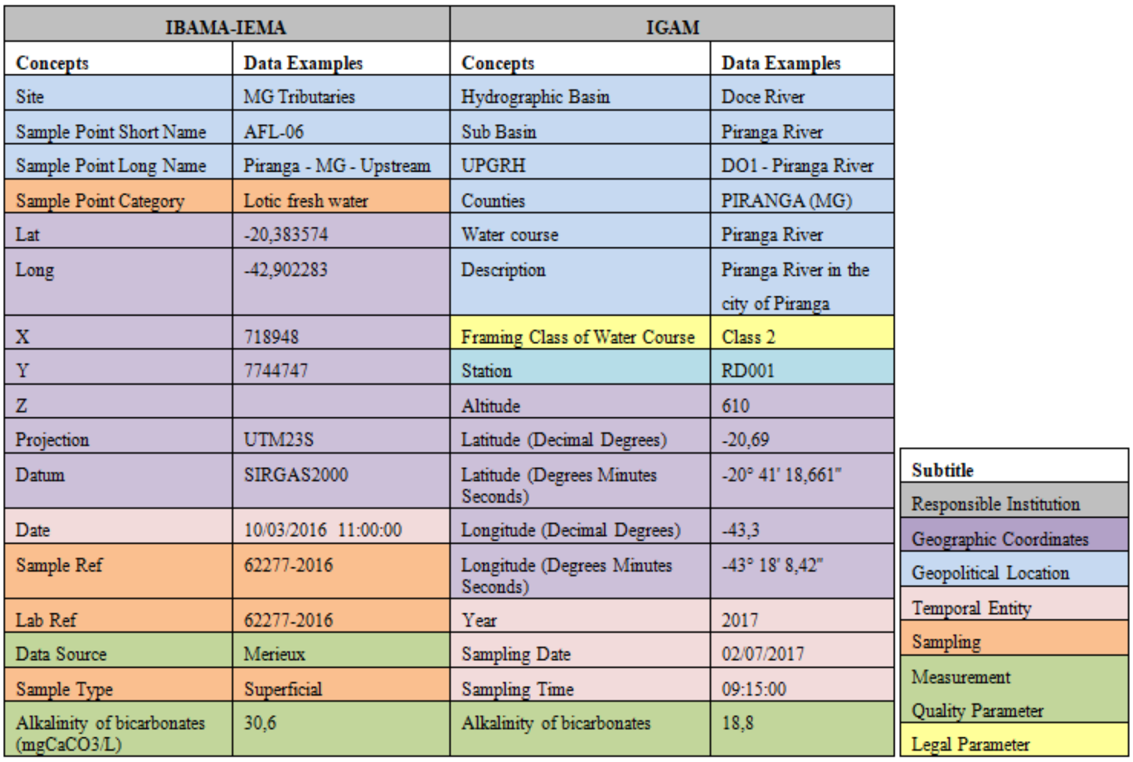
The proposed approach has as starting point the identification and analysis of existent real-world data with the help of domain experts. By analyzing such data, it is possible to identify the problems presented by them, but also the relevant concepts of the domain. In other words, it provides means to define an initial scope of the ontology network to support EQR. Once we have defined the initial scope, we need to search for knowledge sources that address the relevant concepts of the domain so that they can be evaluated regarding their coverage, accuracy and adequacy to the data so that they can be reused in the ontology network. These main activities will be discussed below.

3. Environmental Quality Data Sources Identification and Analysis

In the context of our project, several sources of Brazilian water quality data were analyzed, mainly data sources generated to assess the impacts of the *Mariana* disaster. The analysis of such data highlights the problems presented by them but also helps to identify the key concepts of the domain.

Table 1 shows water quality data from two Brazilian government agencies: data from (IBAMA, 2018) and (IEMA, 2018) in the first two columns and from (IGAM, 2018) in the last two. Several problems can be observed in relation to these datasets, starting with terms used to identify them and their granularity, which vary according the source, showing their heterogeneity. For example, (IBAMA, 2018) and (IEMA, 2018) use “Sample Point Long Name” to identify river, state and location of collect. (IGAM, 2018) uses “Water course” to identify river, “Counties” to identify city and state and “Description” to identify location of sampling.

Table 1. Concepts of Water Quality Used by Different Brazilian Agencies



IBAMA-IEMA		IGAM	
Concepts	**Data Examples**	**Concepts**	**Data Examples**
Site	MG Tributaries	Hydrographic Basin	Doce River
Sample Point Short Name	AFL-06	Sub Basin	Piranga River
Sample Point Long Name	Piranga - MG - Upstream	UPGRH	DO1 - Piranga River
Sample Point Category	Lotic fresh water	Counties	PIRANGA (MG)
Lat	-20,383574	Water course	Piranga River
Long	-42,902283	Description	Piranga River in the city of Piranga
X	718948	Framing Class of Water Course	Class 2
Y	7744747	Station	RD001
Z		Altitude	610
Projection	UTM23S	Latitude (Decimal Degrees)	-20,69
Datum	SIRGAS2000	Latitude (Degrees Minutes Seconds)	-20° 41' 18,661"
Date	10/03/2016 11:00:00	Longitude (Decimal Degrees)	-43,3
Sample Ref	62277-2016	Longitude (Degrees Minutes Seconds)	-43° 18' 8,42"
Lab Ref	62277-2016	Year	2017
Data Source	Merieux	Sampling Date	02/07/2017
Sample Type	Superficial	Sampling Time	09:15:00
Alkalinity of bicarbonates (mgCaCO3/L)	30,6	Alkalinity of bicarbonates	18,8

Subtitle
Responsible Institution
Geographic Coordinates
Geopolitical Location
Temporal Entity
Sampling
Measurement Quality Parameter
Legal Parameter

Another problem observed in the data of Table 1 is the measurement unit of the water quality parameters, which is informed by (IBAMA, 2018) and (IEMA, 2018), but not by (IGAM, 2018). In the case of data from IGAM the identification of units must be done by whoever is interpreting the data, which may cause problems. Also, one can verify the lack of quality of the data and vocabularies employed. For example, “Sample Type” is used with two different purposes. When filled with *Superficial*, this means that the sample material type is surface water. When it is filled with *Daphnia similis*, this means that an ecotoxicological bioassay with the species *Daphnia similis* was performed. This compromises clarity and can only be inferred by domain experts.

Despite the heterogeneity and the data problems of the different agencies, it is possible to identify the relevant concepts that characterize EQR. They are (as grouped in Table 1): responsible institute, geographic coordinates, geopolitical location, temporal entity, sampling (sample: *62277-2016*, sample material type: *Superficial* and sampling procedure), quality parameter (Alkalinity of bicarbonates), measurement (measured values, units of measurement and analytical laboratory: *Merieux*) and legal parameter.

4. Environmental Quality Knowledge Sources Identification and Analysis

Based on the representative concepts of EQR, a search for existing knowledge sources (ontologies, vocabularies and standards) that cover this domain must be performed to enable the reuse of them and avoid the creation of new resources unnecessary. Table 2 shows the relation of the resources found for the concepts identified in section 3.

Table 2. Identified Knowledge Sources

Ontology, Standard or Vocabulary	Description	Key Concepts
om-lite (Cox, 2017)	Ontology for observation features, based on the O&M conceptual model from OGC and ISO 19156.	Feature of Interest, Observed Property, Result, Procedure, Phenomenon Time and Result Time
sam-lite (Cox, 2017)	Ontology for sampling features, based on the O&M conceptual model from OGC and ISO 19156.	Sampled Feature, Shape, Sampling Time, Sampling Method, Sampling Location, Current Location and Size
QUDT (Simons et al., 2013)	Quantities, Units, Dimensions, Data Types ontology.	Units of Measure, Quantity Kinds, Dimensions and Data Types
ChEBI (Degtyarenko et al., 2008)	Chemical Entities of Biological Interest ontology.	ChEBI Name, ChEBI ID, Definition and Synonym
OP (Cox et al., 2014)	Ontology for observable properties which extends QUDT, incorporating some of the requirements identified in the O&M model and its successors.	Scaled Quantity Kind, Quality Kind, Applicable Vocabulary, Property Kind, Feature of Interest, Procedure and Substance or Taxon
SSN (W3C SSN-XG, 2005)	The Semantic Sensor Network ontology describes sensors and their observations, the involved procedures, the studied features of interest, and the observed properties.	Actuatable Property, Actuator, Feature of Interest, Observable Property, Result, Procedure, Sample and Sensor
M-OPL (Barcellos et al., 2014)	Measurement Ontology Pattern Language addresses the measurement core conceptualization. Can be used for building measurement ontologies to several domains.	Measurable Entities, Measures, Measurement Procedures, Measurement Units & Scales and Analysis
Darwin Core (Darwin Core Task Group, 2014)	Body of standards for biodiversity informatics. It provides stable terms and vocabularies for sharing biodiversity data.	Taxon, Identification, Occurrence, Record level, Location, Event and Material Sample
OWL-Time (Cox and Little, 2017)	Ontology that provides a vocabulary for topological relations among instants and intervals and information about durations and temporal position.	Date Time Description, Date Time Interval, Time instant, Time interval, Temporal duration and Temporal entity
Basic Geo (WGS84 lat/long) Vocabulary (Brickley, 2003)	RDF vocabulary that provides a namespace for representing lat(itude) and long(itude), using WGS84 as a reference datum.	Points, Latitude, Longitude and Altitude
OntoBio (Albuquerque et al., 2015)	Biodiversity Ontology.	Ecosystem, Environment, Spatial Location, Collect and Material Entity

These knowledge sources should be analyzed for domain coverage based on the representative concepts of EQR. Table 3 presents the results of this analysis, showing the concepts that are treated by each ontology, vocabulary or standard of Table 2. As can be seen, there are already resources to deal with most of the concepts raised and they will be analyzed for reuse. However, some of them, because they are very specific domains, applicable only at national, state or municipal level, as in the case of legal parameters, are not treated by existing resources and need to be structured to be coupled to the ontology network.

The analysis of Table 3 allows us further to infer the initial modularization of the ontology network. All the concepts of sampling are treated by four of the eleven resources identified. In addition, the measured values and the units of measure of the measurement group are covered, respectively, by six and four of the eleven resources identified. Therefore, they should probably be approached as core ontologies (Ruy et al., 2016).

The other concepts are used to answer questions related to the concepts of sampling and measurement, such as: sampling location, date and time of these activities, measured parameters, between others. So, they should probably be approached as domain ontologies (Ruy et al., 2016).

Table 3. Knowledge Sources and their Domain Coverage

Knowledge Sources Concepts	om-lite	sam-lite	QUDT	ChEBI	OP	SSN	M-OPL	Darwin Core	OWL-Time	Basic Geo	OntoBio
Sampling: *Sample*		x				x		x			x
Sample Material Type		x				x		x			x
Sampling Procedure	x	x				x					x
Measurement: *Measured Values*	x		x		x	x	x	x			
Units of Measurement			x		x		x	x			
Analytical Laboratory							x	x			
Geopolitical Location								x			x
Geographic Coordinates								x		x	x
Temporal Entity								x	x		
Quality Parameters: *Physical*			x		x						
Chemical				x	x						
Biological					x			x			x
Legal Parameters											
Responsible Institution								x			

5. Final Considerations and Future Work

This paper presents the first steps in the building of an ontology network to support EQR. The bottom-up approach of analyzing the real environmental quality data of several institutions was carried out, making it possible to identify the representative concepts of the domain, despite the heterogeneity of the data. Then, a search for knowledge sources that cover these concepts was made. Lastly, the coverage of the concepts by the sources was analysed, making it possible to conclude on resources that can be reused; to identify the concepts that are not treated and, therefore, should be the subject of new ontologies to be developed; and to establish the initial modularization of the ontology network.

As future work, we intend to follow up on the search for knowledge sources in a systematic way so that no relevant resource is disregarded in the elaboration of the ontology network. After, we must complete the identification of the core ontologies and develop them. Then, we need to match the domain ontologies necessary for the modelling of EQR and to develop those whenever their reuse is not possible. Finally, we intend to develop an eScience portal to publicize the original EQR data annotated with systematized metadata. The objective is for data from various sources to be integrated and made available to the community in general.

Acknowledgments

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

References

Albuquerque, A.C.F., Dos Santos, J.L.C., De Castro, A.N. (2015) OntoBio: A biodiversity domain ontology for Amazonian biological collected objects, in: Proc. Annual Hawaii International Conference on System Sciences. pp. 3770–3779.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T. (2000) Gene Ontology: tool for the unification of biology. Nat. Genet. 25, 25.

Barcellos, M.P., Falbo, R.A., Frauches, V.G.V. (2014) Towards a measurement ontology pattern language, in: Proc. 1st Joint Workshop ONTO.COM/ODISE, CEUR Workshop Proceedings, Vol,. 1301.

Brickley, D. (2003). Basic Geo (WGS84 lat/long) Vocabulary. W3C Semantic Web Interest Group, W3C.

Cox, S.J.D. (2017). Ontology for observations and sampling features, with alignments to existing models. Semantic Web 8, 453–470.

Cox, S.J.D., Little, C. (eds.) (2017) Time Ontology in OWL. W3C Recommendation.

Cox, S.J.D., Simons, B.A., Yu, J. (2014) A Harmonized Vocabulary For Water Quality, in: HIC2014 - 11th International Conference on Hydroinformatics. pp. 1454–1459.

Degtyarenko, K., De Matos, P., Ennis, M., Hastings, J., Zbinden, M., Mcnaught, A., Alcántara, R., Darsow, M., Guedj, M., Ashburner, M. (2008) ChEBI: A database and ontology for chemical entities of biological interest. Nucleic Acids Res. 36.

FORCE11 (2016) The FAIR Data Principles [WWW Document]. La Jolla, CA FORCE11. URL https://www.force11.org/group/fairgroup/fairprinciples

Darwin Core Task Group (2014) Darwin Core Terms: A complete historical record [WWW Document]. TDWG. URL http://rs.tdwg.org/dwc/terms/history/

W3C SSN-XG (2005) Semantic Sensor Network Ontology [WWW Document]. online url https//www.w3.org/2005/Incubator/ssn/ssnx/ssn.

IBAMA [WWW Document], 2018. URL http://www.ibama.gov.br/

IEMA [WWW Document], 2018. URL https://iema.es.gov.br/

IGAM [WWW Document], 2018. URL http://www.igam.mg.gov.br/

Ruy, F.B., Falbo, R.A., Barcellos, M.P., Costa, S.D. and Guizzardi G. (2016) SEON: A software engineering ontology network, in: European Knowledge Acquisition Workshop, pp. 527–542.

Simons, B.A., Yu, J., Cox, S.J.D. (2013) Defining a water quality vocabulary using QUDT and ChEBI, in: Proc. 20th Intl. Congress on Modelling and Simulation. Modelling and Simulation Society of Australia and New Zealand, pp. 2548–2554.

Soares, A. (2009) Towards ontology-driven information systems: Guidelines to the creation of new methodologies to build ontologies, PhD dissertation, Penn. State Univ.

Suárez-Figueroa, M., Gómez-Pérez, A., Fernández-López, M. (2012) The NeOn Methodology for Ontology Engineering, in: Suárez-Figueroa, et al. (Eds.), Ontology Engineering in a Networked World SE - 2. Springer Berlin Heidelberg, pp. 9–34.