Formal Ontology in Information Systems
M. Donnelly and G. Guizzardi (Eds.)
IOS Press, 2012

doi:10.3233/978-1-61499-084-0-73

Towards Making Explicit the Ontological Commitment of a Database Schema on the Geological Domain

Alda Maria Ferreira Rosa da Silva[a,b] and Maria Cláudia Cavalcanti[a,1]
[a] *Department of Computer Engineering, Instituto Militar de Engenharia (IME), Brazil*
[b] *Geoprocessing Division, Companhia de Pesquisa de Recursos Minerais (CPRM), Brazil.*

Abstract. Recently, the demand for data integration on the Geological domain has increased. Most approaches for database schema integration are strongly based on structure and syntax, and present limitations. Semantic resources, such as ontologies, have been used to reach better results for data integration. However, these approaches assume there are already local ontologies that represent the databases involved in the interoperation. Furthermore, the creation of an ontology based on the database logical schema is not an easy task. On the other hand, it had been proposed that the association of a top-level ontology to a database conceptual schema may lead to better interoperability results. The idea is to make explicit the ontological commitment of each representation, and thus facilitate their integration. This work presents a case study on the Geological domain, which aimed at making explicit the ontological commitment of a database conceptual schema, in order to further improve data interoperability. The main contribution of this work is that it covers the whole process. It starts from the database logical schema, applies a set of reverse engineering techniques to create a preliminary database conceptual schema, and then uses a top-level ontology as a way of making explicit its ontological commitment. A detailed description is provided on how each step was taken, serving as a roadmap for others that may need to go through a similar process on the geological domain or on other domains.

Keywords. Interoperability, top-level ontology, Lithostratigraphy

1. Introduction

Data integration is an old and widely-explored problem. More recently, the demand for data integration on the Geological domain has increased. Nowadays, geological information is on our day-by-day tasks. GPS mechanisms may be used to request information about, for instance, the soil of some plot of land near the beach or near some mountain, to obtain technical support and investigate if it is safe to build or cultivate vegetables on that area. The current scenario of geological data is not very promising due to: (i) there are lots of non-structured data (images, reports, etc.), which are difficult to manipulate; (ii) structured and semi-structured data from different and

[1] Corresponding Author: Maria Cláudia Cavalcanti, Seção de Engenharia de Computação, Instituto Militar de Engenharia, Praça General Tiburcio, 80, Praia Vermelha – Urca, Rio de Janeiro, RJ, 22290-270 Brazil; E-mail: yoko@ime.eb.br.

heterogeneous sources, which are difficult to integrate; (iii) lack of an efficient interoperability policy.

Geological data come from abstractions of natural phenomena of the real world, which means they are an interpretation of reality. According to Kent [38] "the resemblance between the extracted ideas and the ideas in the observer's mind … depends heavily on the participants' common understanding". Therefore, there may be different interpretations of the same natural phenomena. Moreover, the nomenclature used to represent such phenomena and their relations is frequently different, meaning these data are organized differently. Some initiatives, such as the OpenGis [33] and the INDE [25] here in Brazil, contribute to the establishment of an interoperability policy. OpenGis proposes a set of technologies for geological data interoperability. For instance, they propose controlled vocabularies, taxonomies, file formats, etc. However, these initiatives are top-down approaches, and their adoption has just started in Brazil. This adoption may take many years because there are many legacy systems in this domain.

This adoption difficulty is present in many other areas, where top-down approaches had also been proposed. On the other hand, there are also bottom-up approaches, which may lead to a more efficient and local adoption. Over the years, researchers of the database area have been investigating and proposing solutions for local schema integration. The schema integration traditional approach [29] proposes the use of a central global schema, to which data schemas would be mapped to. A more flexible and decentralized approach is the mediation architecture [22], in which modules called wrappers would be responsible for data transformations between different schemas. These approaches were strongly based on the syntax of the schemas, and present limitations for not considering the meaning behind each entity represented in such schemas.

Some recent approaches [10][13] use semantic resources, such as ontologies to provide more promising solutions for the interoperability problem. Calvanese *et al.* [10] propose a two-level architecture to solve such problem. At the lower level there are Local Ontologies (LO). Each LO describes each local data sources. At the upper level there is the Domain Ontology (DO), which contains the basic terms of a domain. LO-DO mappings are used to provide data interoperability. Similarly, in [13] the authors propose a model for specifying an ontology vocabulary matching, considering their schemas as local ontologies.

Although these approaches facilitate the interoperation between Local Ontologies, they do not explain how to create an ontology based on a database logical schema. This is a hard task, and demands to understand and capture the meaning behind each concept. Even if the system provides a schema at a higher level of abstraction, such as a conceptual schema, according to [15], conceptual schemas and ontologies belong to different epistemic levels, have different objects and are created with different objectives, and thus cannot be taken as equivalents.

In their work, Guizzardi and Wagner [19] state that the intended meaning embedded in the entities of either a conceptual schema or an ontology representation should be made explicit through the association to a system of meta-level categories, or a top-level ontology, named UFO. Another previous and similar work proposes the VERONTO technique [28]. This association is also known as to establish the ontological commitment. An interesting work on this direction shows that the establishment of such ontological commitment generates better ontology alignment results [37]. In this work, two domain ontologies are aligned after enrichment by their

integration to a top-level ontology. However, an important question still remains: how to reach such high level conceptualization if all you have so far is a set of database logical schemas?

This work focus on this problem and combines a set of reverse engineering techniques and the use of top-level ontologies as a way to make explicit the ontological commitment of a conceptual schema, in order to facilitate the interoperability of the original database. We describe a case study on the geological domain that started with the redemption of the documentation of a database logical schema, which was lost or inexistent. Some reverse engineering techniques [11][35] helped us on the first steps of the process. However, these techniques do not reach the semantics of the data. Therefore, in order to make explicit the semantics of the conceptual schema and complete the process, we added the application of the OntoClean methodology [31] and OntoUML meta-categorization [17]. The main contribution of this effort is to identify a set of actions (guidelines) that could be applied to other database schemas.

This paper is organized as follows. The next section presents some of the main concepts in the domain of the Lithostratigraphy. Section 3 defines some of the basic concepts about Ontologies that will be used throughout this paper. Section 4 describes the case study, detailing the steps taken towards creating the ontologically committed conceptual schema on the Lithostratigraphy domain. Finally, we discuss related and future work and conclude the paper.

2. Lithostratigraphy Domain

In order to use ontological formalisms, first it is necessary to identify and organize well founded concepts within a specific domain. The domain of stratigraphy deals with the shape, arrangement and rock layers that form the Earth's crust [36]. This domain may be ruled according to several distinct concepts. In this work, the scope chosen is the Lithostratigraphy, which is the area that describes the arrangement of strata (rock layers) in a certain area, considering their lithology, mineralogy, grain size and stratigraphy. The Lithostratigraphy is limited to what is arising on the surface (specifically, the upper crust), and therefore, involves the study of the formation of the Earth and its rocks (aggregates of one or more minerals). Each *rock* can be genetically classified (ie, according to their origin) as: *Igneous* or *Magmatic* formed from the cooling of magma, *Sedimentary*, formed from fragments of other rocks, and *Metamorphic* - formed from the transformation of sedimentary or igneous rocks, through processes involving changes in pressure and temperature.

When performing the characterization of a cluster of rocks through their lithology (description of the rocks in outcrop or hand sample, based on various characteristics),or chronology (age of rocks), the geologist creates a *Stratigraphic Unit*. These units are classified into: *Lithostratigraphic Units*, formed by rocks of the same lithological characteristics which form layers or strata; *Lithodemic Units*, formed by rocks that are not formed in layers, but show well-defined contacts with other rocks. A *Lithodemic Unit*, may be sub-classified as a *Complex*, featuring a set of rocks that result from the meeting or mixture of two or more genetic rocks, which cannot be mapped separately, grouping the informal units and their lithofacies [24].

To date each *stratigraphic unit*, it is associated to a *Geochronological Unit*. Each *geochronological unit* corresponds to a division of the geologic time scale. When analyzing a *stratigraphic unit,* its association to a *geochronological unit* determines the

age of the Earth in which it emerged. This division of the geologic time scale is organized in hierarchical levels. The Eon units (Hadean, Archean, Proterozoic and Phanerozoic), are the largest geochronological units. The Era units (Paleozoic, Mesozoic and Cenozoic) are subdivisions of the Phanerozoic Eon. The Period units are the subdivisions of the Era units. Finally, the Epoch units are subdivisions of Period units, existing only within the Cenozoic Era.

3. Ontologies and basic concepts

According to Calvanese *et al.* [10] there are ontologies at different abstraction levels. While a domain ontology (DO) represents the conceptualization of a specific domain, local ontologies (LO) have been used to formally describe the semantics of a specific data source and to make its content explicit [23]. However, while a DO is built out of a set of specialists perception of what is important in a given domain, an LO provides a very specific view of reality, and as stated in [38], the representation of an information of the real world as a data structure results from a process of communication among people, and depends heavily on the participants' common understanding of it. There is a lot of misunderstanding on what is "one thing" in the real world, and such misunderstanding may be amplified through data structures, as they try to represent the original observer's idea.

From a data integration perspective, ontologies provide a possible approach to address the problem of semantic heterogeneity [13]. Through a formal ontological language, one can make explicit all concepts behind data structures. Therefore, local ontologies can be extracted from data schemas, and they could be used, instead of schemas to facilitate the identification of correspondences between the data sources behind those schemas. However, aligning LOs is not an easy task. As we said before, an LO provides a very specific view of a reality, and thus, ambiguities and misunderstandings may lead to false agreements, i.e., these ontologies may seem to have a shared view of reality, but they reflect just a small group conceptualization. For instance, an *annotation* is a product for some, and a process for others.

In a previous and important ontology-level classification [30], the author includes a top-level ontology, defined as an ontology that describes very general concepts such as space, time, matter, object, event, action, etc, which are independent of any particular domain. Guizzardi [17] states that top-ontologies have well-founded philosophical concepts, and proposed their use to make explicit the ontological commitment of a specific conceptual schema, and help prevent the false agreement on further model integration.

The Unified Foundational Ontology (UFO) is a top-level ontology that is divided in three parts: A, B and C. The UFO-A foundational ontology, proposed in [17], is an ontology of endurants (objects), which addresses issues such as: (i) the general notions of types and their instances; (ii) objects, their intrinsic properties and property-value spaces; (iii) the relation between identity and classification; (iii) distinctions among sorts of types (e.g., kinds, roles, phases, mixins) and their admissible relations; (iv) distinctions among sorts of relational properties; (v) Part-whole relations.

An UFO fragment (mainly UFO-A elements/meta-categories) is presented in Figure 1. According to Guizzardi, endurants (continuants) are entities that exist in time while keeping their identity. Endurants are said to be wholly (it and its parts) present whenever they are present. Examples of endurants are a person or an amount of sand.

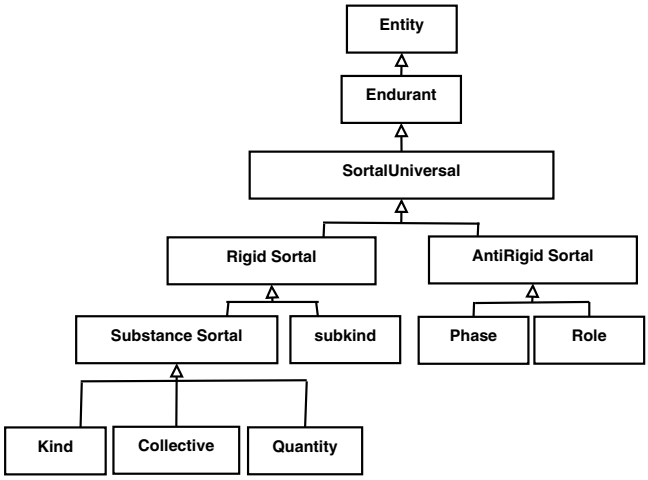

Figure 1: UFO Fragment - Adapted from [5]

As shown in Figure 1, sortals are endurant entities that are subcategorized as kind, subkind, quantity, phase, role and collective. Kinds and subkinds represent instances that can be identified and individualized. For instance, person, football player and child are examples of kind, role and phase, respectively. Quantities or amounts of matter are typically referred to by means of mass nouns [19]. In order to enable references to general terms which are not count nouns, they first must be nominalized. A nominalization of a mass noun promotes the shift to the category of count nouns. For instance, "the water" becomes "the water in the bathtub". Other examples are: a lump of clay, a cube of sugar, a liter of water, a piece of gold, a pile of sand). In [20], the authors summarize by saying that a "quantity" represents portions of amounts of matter.

Each UFO-A concept is characterized by a set of principles (meta-properties). For instance, kinds and subkinds are endurant entities that can provide the principle of identity and individuation (unity) for its instances; while quantity entities do not satisfy the principle of unity. Therefore, the identification of such principles for each concept of a conceptual schema is a way to identify to which UFO meta-category it belongs to.

Guarino and Welty [31] proposed the OntoClean methodology as a guide to validate concepts, properties and taxonomic relations of conceptual schemas. This methodology may be useful to make explicit certain aspects of the ontological commitment of specific conceptual schemas, and allows the user to investigate the correctness of each taxonomic relation. It is based on the philosophical notions of essence, identity, unity and dependence. The idea is to use these notions as a theoretical framework of meta-properties, to analyze each concept and related properties of a given schema, and contrast them with such notions.

The notion of essence is related to rigidity, i.e., if a concept represents the essence of an instance then this instance MUST hold for it. In other words, a concept is said to be Rigid (+R), if for all instances of that concept, they cannot stop being an instance of this concept in any possible world. But, if there is a single instance that does not hold for it, than it is said to be NOT Rigid (-R). And finally, there is the notion of anti-rigidity (~R), for a concept for which all instances of it must possibly not be instances of it.

Identity refers to the problem of being able to recognize individual entities (instances) in the world as being the same (or different). Identity criteria (IC) are

conditions used to determine equality (sufficient conditions) and that are entailed by equality (necessary conditions). For instance, if two persons a and b, have the same fingerprints than they are the same person. Therefore having-the-same-fingerprints is a sufficient condition for identifying a person. On the other hand, if two persons a and b are the same person then they have the same fingerprints (necessary condition). A more formal definition of such notion can be found in [32]. For this paper, we have also to distinguish between the notion of carrying or not identity (+I/-I), and supplying or not identity (+O/-O). If a concept C_1 carries identity conditions then it supplies it to any subconcept C_2. Therefore, C_1 is classified as +O while C_2 is classified as +I. In addition, Raban and Garner [34] say that if a concept has its own identity (+O) it also has identity (+I), and, conversely, not having identity (-I) rules out its own identity, and therefore, implies lack of own identity (-O).

The notion of unity refers to being able to recognize which are the parts (and which are not) that form an individual (instance), how and in which conditions they are connected as a whole. This set of conditions is known as Unity Constraints or Unity Conditions (UC). If it is not possible to recognize the parts of an instance of a concept, then this concept is said to have no unity (-U). On the other hand, a concept is said to have unity (+U) if all its instances are intrinsic wholes. An interesting example of –U is the "water" concept, whose instances are "amounts of water", and for which it is not possible to define a UC. And finally, there is the notion of anti-unity (~U), for a concept for which ALL instances cannot be identified by their parts and limits.

The notion of dependence is related to external dependence [32]. It says that a concept C_1 is externally dependent (+D) on a concept C_2 if, for all its instances, necessarily some instance of C_2 must exist, which is neither a part nor a constituent of C_1. The classical example is that of a parent that should be related to a child, i.e., an instance of a parent must be related to an instance of a child. If there is no external dependency then a concept C_1 is said to be not dependent (–D).

4. CPRM case study

This section presents a case study on the Geological domain. It describes how an ontologically committed conceptual schema was created, from the LITO database. LITO is one of the databases in use at CPRM (Companhia de Pesquisa de Recursos Minerais), the Brazilian Company for Mineral Resources Research, which stores Litostratigraphic data. After describing the LITO database briefly, a set of steps are described. The first step included a set of actions taken to analyze LITO database, in order to identify its core and to create a preliminary conceptual schema. Then, a second step involved the application of the OntoClean Methodology, explained previously, i.e., OntoClean meta-properties were assigned to each concept of the LITO conceptual schema. Finally, a third step consisted of the meta-categorization of the LITO conceptual schema, according to the meta-properties and the UFO-A top-ontology, generating the LITO ontologically committed conceptual schema.

4.1. LITO Database

The set of databases are known as Geobank [26], and it was developed by the DIGEOP (Geoprocessing Division) team, with the support of many specialists in the company.

Nowadays, LITO is used by many CPRM departments, and it is remotely accessed by universities, city halls, and some South American countries.

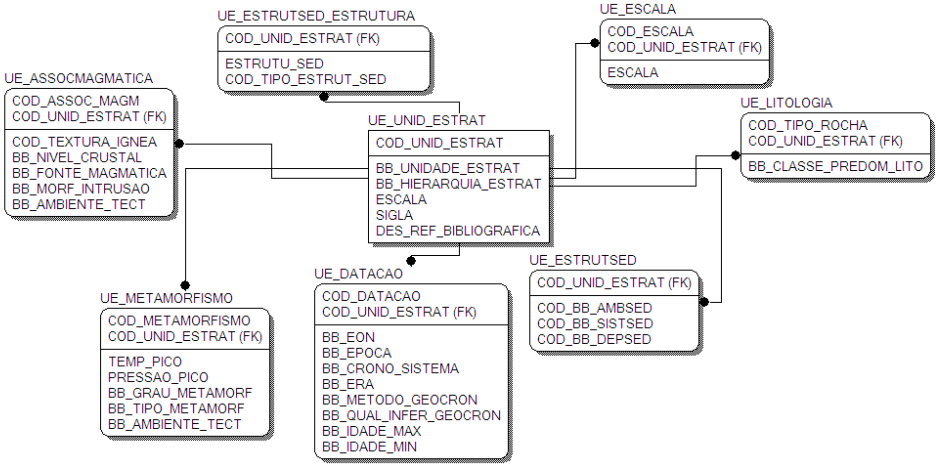

Figure 2: LITO Logical Database Schema Fragment

According to LITO documentation [9], LITO database holds a set of registries, named Lithostratigraphic Units (UE). These units are associated to one or more cartographic scales, and are identified uniquely by an acronym/symbol ("letra símbolo"). Each LITO database unit corresponds to a map polygon stored at Geobank.

Figure 2 shows a fragment of the LITO logical database schema, and its main tables. A selection of tables and attributes was done based on what data would probably be exchanged with other systems. A mini data dictionary, with table and attribute descriptions is presented next.

- `UE_UNID_ESTRAT` – Stratigraphic unit table:
  `bb_unidade_estrat` – name of the stratigraphic unit;
  `bb_hierarquia_estrat` – hierarchy of the stratigraphic unit;
  `escala` – scale;
  `sigla` – acronym;
  `des_ref_bibliografica` – bibliographic reference.
- `UE_ESCALA` – Scale table:
  `escala` – scale.
- `UE_datacao` – related stone age table:
  `bb_crono_sistema` – geo-chronological system use;
  `[bb_eon | bb_era | bb_epoca ]` – eon, era or epoch;
  `bb_metodo_geocron` – geochronology method used;
  `bb_qual_infer_geocron` – inference quality;
  `idade_maxima` – maximun age;
  `idade_minima` – minimal age.
- `UE_assocmagmatica` – igneous or magmatic rocks table:
  `cod_assoc_magm` – magmatic association rock;
  `cod_textura_ignea` – igneous rock texture;
  `bb_nivel_crustal` – crustal level of ígnea rock;
  `bb_fonte_magmatica` – magma source that originated the rock;

`bb_morf_intrusão` – intrusion morphology of the ígnea rock;
`bb_ambiente_tect` – tectonic environment.

- `UE_metamorfismo` – metamorphic rock table:
  `cod_metamorfismo` – id for the metamorphic rock;
  `temp_pico` – metamorphism temperature;
  `pressão_pico` – metamorphism pressure peak;
  `bb_cod_grau_metamorf` – metamorphism grade;
  `bb_tipo_metamorf` – metamorphism type;
  `bb_ambiente_tect` – tectonic environment.
- `UE_estrutsed` – sediment table:
  `cod_bb_ambsed` – sedimentation environment;
  `cod_bb_sistsed` – sedimentation system;
  `cod_bb_depsed` – sedimentation deposit.
- `UE_estrutsed_estrut` – sedimentar structures table:
  `estrut_sed` – sedimentar structures;
  `cod_tipo_estrut_sed` – sedimentar structure type.
- `UE_litologia` – rocks table:
  `cod_tipo_rocha` – rock type;
  `bb_classe_predom_lito` – predominance of a rock.

4.2. LITO Database schema analysis

The analysis of the LITO logical schema aimed at capturing the concepts of its domain. The first step consisted of generating an automatic version of the conceptual schema of LITO, based on its logical schema. To do this, we counted on the Oracle schema extraction tool. Inspired by the reverse engineering literature [6][8][6][35][11], we adopted the following steps: (i) main tables identification; (ii) hierarchies identification; (iii) relevant attribute selection; (iv) UML preliminary schema generation. Through these steps it was possible to generate a schema representation at a higher level of abstraction (conceptual schema), which facilitates the understanding of the relations of the original logical schema and their relationships.

In order to identify the main tables (step i), each of them was analyzed with respect to one of the already known main concepts of the domain of the database, such as stratigraphy, stratigraphic units and their characterization, chronology (Earth age), scales, rocks and their specializations, sedimentary environment, etc. Another criterion used was the investigation of instances of these tables to confirm the meaning behind some attribute and/or table. To support this task CPRM technical documentation (manuals) about the system procedures and database were used. Additionally, and more importantly, it was possible to count on some domain and system specialists who guided the documentation and database schema analysis, as well as, indicated specialized literature. At the end of this step we already were able to generate a draft of the LITO conceptual schema, with the main concepts that represent the selected main tables and their relationships.

With respect to the identification of hierarchies (step ii), all association and subsumption relationships between concepts were analyzed. Relationships between two tables t_1 and t_2 may represent some kind of classification of a tuple of t_1 by a tuple of t_2. Just looking at the schema table names, it is sometimes possible to identify which is the

classifier and which is the classified. In relationships with cardinality (1:n), the classified concept is more easily identified as the one which participates, at most, one time at the relationship. However, for the other cardinalities ((1:1), (n:m)) this identification demands a deeper look at the attributes and tuples of each involved table, in order to identify not only which is the classifier table, but also if its tuples hide some hierarchy. If this is the case, the extracted conceptual schema will include such classifier tables as concepts representing explicit hierarchies.

In the LITO conceptual schema (Figure 3), Rock and Stratigraphic Hierarchy concepts, and their corresponding hierarchies (sub-concepts), were identified from classifier tables such as `UE_metamorfismo`, `UE_estrutsed`, `UE_assocmagmatico`, and `UE_estrutsed_estrutura` and their relationships to table `UE_unid_estrat`, all tables of the LITO logical schema (Figure 2).

Step iii started after identifying the main tables and hierarchies. It involved the analysis of each selected table attribute, aiming at discarding those that were not relevant for interoperation. Identifier attributes (candidate keys) were maintained, together with those attributes that add descriptive and complementary information to each tuple of the table. Superfluous and administrative (e.g. auditing) attributes were discarded. This analysis and selection was performed with the help of CPRM specialists and system documentation. At the end of this step we could detail the LITO conceptual schema by including the attributes of each concept.

4.3. Application of the OntoClean Methodology

With this first version of LITO conceptual schema in hand it was possible to step towards making explicit its ontological commitment. Each concept of the LITO conceptual schema was then analyzed in accordance to the OntoClean meta-properties, described previously: Identity, Rigidity, Unity and Dependency. The following descriptions explain how each concept was classified according to these meta-properties. Table 1 presents a summary of this classification.

Rock (+O, +R, ~U, -D): A rock is an aggregate of one or more individual minerals, and each rock is characterized according to such minerals. This characterization is considered stable, as was done by studying the change of the Earth in billions of years. In our concept of time, this characterization will not change. Thus the rock concept is classified as rigid (+R), meaning a rock instance will always be a rock throughout its existence. A rock provides identity (+O) because it is possible to identify its instances, through the percentage of the essential minerals that constitutes a rock (dominant percentage defines which rock it is), and through the way it was formed (which defines the rock texture, size, shape and arrangement of minerals). These are the conditions that constitute the identity criteria of the rock concept. If it provides identity, then, by definition, it also holds it (+I). A rock is understood as an amount of matter, because it is not possible to identify its parts and boundaries. Therefore, the rock concept has no unity criteria (~U). Finally, a rock is not existentially dependent on any other concept of the domain (-D).

For all subtypes of rock - Igneous, Sedimentary, Metamorphic - and their subtypes – (i) Plutonic, Volcanic and Subvolcanic; (ii) Clastic, Chemical and Biogenic; (iii) Regional, Contact, Dynamics, Impact and Hydrothermal, we extend the classification used for the rock concept, i.e., they are rigid (+R), have no unity criteria (~U), and are not dependent on any other concept of the domain (-D). Each rock subtype can be identified (+I). For instance, the igneous rock is identified because it is formed through

the cooling of the magma, which gives it a different texture, and mineralogical constitution. Some subtypes of rock add some other conditions for its identification criteria. For instance, the sedimentary rock is also identified by its texture, structure (faces), stiffness, viscosity, density, among others.

Scale (+I, +R, +U, -D): According to the dictionary a scale is "an indication of the relationship between the distances on a map and the corresponding actual distances". In other words, it indicates the ratio used for the graphical representation of a geographical space. The scale concept is classified as rigid (+R), as a scale instance will always be a scale throughout its existence. The scale concept does not provide identity (-O), it has an identity criteria (+I) formed by the combination of the numbers that constitute ratio. The scale is a ratio for which its parts are known and well defined, and thus it has unity criteria (+U). Finally, it is not existentially dependent on any other concept of the domain (-D).

Stratigraphic unit (+I, +R, +U +D): A stratigraphic unit is the characterization of a rock cluster. The stratigraphic unit concept is classified as rigid (+R), as its instances will always be stratigraphic units throughout their existence. It does not provide identity (-O), but it does have an identity criteria (+I). Every stratigraphic unit is identified through the set of rocks that it describes, and their lithology and chronology. Alternatively, it may be uniquely identified by an acronym/symbol ("letra símbolo"). A stratigraphic unit describes a geographically restricted area of rock layers, not necessarily contiguous. Thus, if it is possible for each instance to delimit its boundary and/or parts, then it has a unity criteria (+U). Because each stratigraphic unit can vary in terms of the scale it uses (decrease or increase the scale), it is existentially dependent (+ D) on the scale concept. Also, a stratigraphic unit makes no sense if not associated to, at least, an instance of a rock, which it describes.

Stratigraphic Hierarchy (+O, +R, -U, -D): A stratigraphic hierarchy is the nomenclature used for the classification of a stratigraphic unit. According to this definition we can classify the stratigraphic hierarchy as rigid (+R) and non-dependent (-D) concept. With respect to the recognition of its parts and borders, it does not provide a unity criteria (-U). On the other hand, it is possible to identify its instances (+I). The unity criteria (+I) were defined by geologists the CPRM Commission's Internal, Department of Geology (more precisely the DIGEOB) [24]. However, as the stratigraphic hierarchy concept is a generic concept and provides its identification criteria to other concepts, then it is classified as an identity provider (+O). The stratigraphic hierarchy is the generic concept of two other sub-concepts that are used to classify the stratigraphic units: (i) lithostratigraphic (+I, +R, -U, -D): supergroup, group, subgroup, formation, member, bed/flow; (ii) lithodemic (+O, +R, -U, -D): supersuite, suite, subgroup, body, facies. The lithodemic concept is an identity provider (+O) because it can yet be specialized in another sub-concept called complex: complex, lithofacies, unit, subunit, zone.

Geochronological Unit (+O, +R, +U, -D): A geochronological unit is defined as a geological time interval. This concept can be classified as rigid (+R). With respect to the identity notion, a time interval is identified by the combination of its first and last instants. Each geochronological unit is classified according to its time interval size, and may be divided into smaller time intervals. All these characteristics define its identity criteria. Therefore, the generic concept (geochronological unit) holds and provides identity (+O), while its sub-concepts (Eon, Era, Period, and Epoch) just hold it (+I). An Eon is the largest geochronological unit. Eras are subdivisions of the Phanerozoic Eon; Periods are subdivisions of the Eras, and Epochs are subdivisions of

the Cenozoic Period. Once parts and boundaries of each geochronological unit are clearly defined, it is then classified as providing a unity criteria (+U). With respect to dependency, an Eon is non-dependent (-D), but each subdivision (Era, Period and Epoch) is dependent (+D) of the larger time interval in which it takes part. For example, each Era instance is dependent on the existence of the Phanerozoic Eon.

In this step we also applied the guidelines suggested by Welty [7] to validate the meta-properties classification. For instance, while classifying the stratigraphic unit, stratigraphic hierarchy, lithostratigraphic, lithodemic and complex concepts, we violated the dependence constraint where +D (dependent) can't subsume –D (independent). This error led us to review such concepts and the associated meta-properties.

Table 1. Classification of the LITO concepts according to the OntoClean meta-properties.

Concept	Identity	Rigidity	Unity	Dependence
Rock	+O	+R	~U	-D
Igneous Rock	+O	+R	~U	-D
Sedimentary Rock	+O	+R	~U	-D
Metamorphic Rock	+O	+R	~U	-D
Scale	+I	+R	+U	-D
Stratigraphic Unit	+I	+R	+U	+D
Stratigraphic Hierarchy	+O	+R	-U	-D
Lithostratigraphic	+I	+R	-U	-D
Lithodemic	+O	+R	-U	-D
Complex	+I	+R	-U	-D
Geochronological Unit	+O	+R	+U	-D
Eon	+I	+R	+U	-D
Era	+I	+R	+U	+D
Period	+I	+R	+U	+D
Epoch	+I	+R	+U	+D

4.4. Meta-categorization of LITO conceptual schema according to UFO-A

In order to identify the UFO-A meta-category for each concept in the LITO conceptual schema, we used the OntoUML tool [2][3]. This task was facilitated because we had already identified the Ontoclean meta-properties (Table 1). Thus, based on [31], and on OntoUML proposal [17], it was possible to meta-categorize each concept (<<stereotypes>>), as shown in Figure 3. For instance, according to OntoClean [31], the Rock concept categorized as (+O, +R, ~U, -D), corresponds to a sortal meta-category, and according to OntoUML, it corresponds to a more specific meta-category, “quantity” (amount of matter, as it provides no unity (~U).

The use of the OntoUML tool was important because we could validate our meta-categorization, once the tool embeds meta-categorization rules such as “a kind could not be a subsumption of a role”. If a violation like this is found, the conceptual schema should be reviewed. More specifically, each concept involved in the violation should be reviewed, in the light of the OntoClean meta-properties already identified, and the UFO-A meta-categories.

In this case-study an example of such violation occurred while meta-categorizing the rock concept and its specializations (igneous, sedimentary and metamorphic) as *quantity*. The OntoUml tool displays the following error message: *"A class stereotyped as «collective», «kind» or «quantity» (the substance sortal classes) cannot have as a supertype a class stereotyped as «kind», «subkind», «quantity» or «collective» (the*

rigid sortal classes), because a substance sortal has its own principle of identity and cannot inherit another principle of identity from other rigid sortal", indicating that the igneous rock concept, could not be a *quantity*. Therefore, it was meta-categorized as *subkind*.

5. Discussion and Related work

Some reverse engineering techniques [11][35] focus on extracting a logical or conceptual database schema diagram based on the data instances. However, these techniques do not reach the semantics of the data. On the other hand, there are works [14][27] that propose an ontology extraction based on some data schema. However, besides considering an existing documentation, such approaches aim to create a domain ontology, and do not focus on making explicit its ontological commitment.

In addition, the top-ontology literature is not rich in providing generic guidelines for how to make explicit such ontological commitment. Nevertheless, some case studies helped us in the work with the LITO conceptual schema, such as: [21] in the software engineering domain; [5] which creates an well-founded ontology in the biomedical domain; [20] which describes a study in the field of Oil and Gas, and analyzes the OWL and OntoUML representations differences; [16] which provides ontological interpretation and modeling guidelines in the representation of types whose instances are quantities (amounts of matter, masses), besides, it analyzes different alternatives for the adequate representation of quantities; and [7] which uses OWL reasoner to check the OntoClean constraints on the taxonomy.

On the geological domain we could count on Lorenzatti work [1] which creates a domain ontology closely related to the LITO database domain. He discusses geological concepts in the light of the Ontoclean meta-properties and UFO-A meta-categories. Another work [12] in this domain uses Ontoclean metaproperties to verify the taxonomic relationships of a Geographical Database Schema.

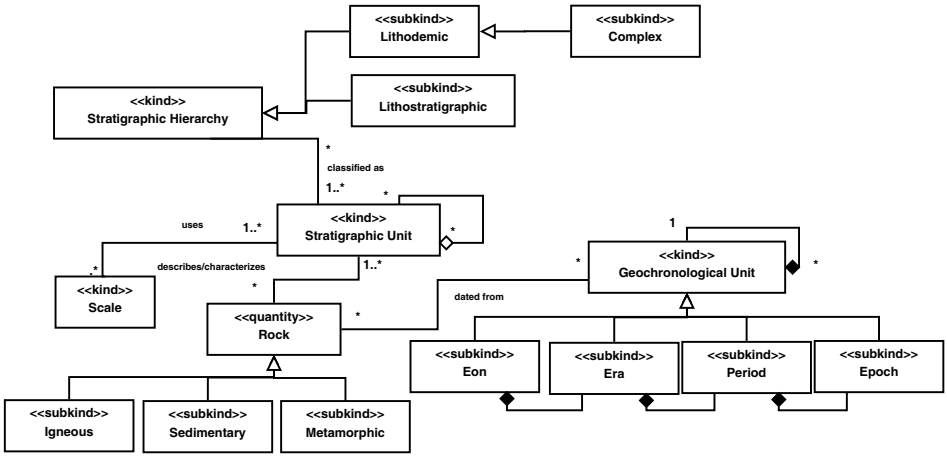

Figure 3: LITO Conceptual Schema meta-categorized at OntoUML tool

6. CONCLUSION

This work presented the CPRM case study, which aimed at creating an ontologically committed conceptual schema, from an existing and poorly-documented database. A detailed description was provided on how each step was taken, serving as a roadmap for others that may need to go through a similar process on geological or other domains. The CPRM case study was the basis to gather and organize actions and guidelines to help such process. A more detailed description of such guidelines can be found in [4].

We are now performing and evaluating ontology alignments using the LITO conceptual schema as the source ontology. We adopted the procedure indicated in [37], and transformed the LITO conceptual schema into an ontology representation using OWL. Additionally, we included in such representation OntoUML meta-categories, subsuming them with LITO concepts, according to the meta-categorization already done. By doing this, we hope to reach better ontology alignments results. Future work includes extending the LITO database table selection and enlarging the final LITO conceptual schema.

Acknowledgements

The authors would like to thank specialists of the Department of Geology, Geoprocessing Division and of the Department of International Topics, at the Brazilian Company for Mineral Resources Research (CPRM) for their collaboration with this work. In addition, they also thank CNPq (309307/2009-0; 486157/2011-3) and FAPERJ (E-26/111.147/2011) for partially funding their research projects.

References

[1] A. Lorenzatti, “Ontologia para Domínio Imaginísticos Combinando Primitivas Textuais e Pictórias”. M.Sc. Dissertation (Mestrado em Ciência da Computação), Univ. Fed. Rio Grande do Sul, Brazil, 2009.

[2] A.B. Benevides, G. Guizzardi, “A Model-Based Tool for Conceptual Modeling and Domain Ontology Engineering in OntoUML”, *Proc. of 11th International Conference on Enterprise Information Systems (ICEIS 2009), Lecture Notes in Business Information Processing* **24** (2009), 528-537, Springer-Verlag.

[3] A.B. Benevides, G. Guizzardi, B.F.B. Braga, J.P.A. Almeida, “Assessing Modal Aspects of OntoUML Conceptual Models in Alloy”, *Proceedings of ER Workshops, Lecture Notes in Computer Science* **5833** (2009), 55-64, Springer.

[4] A.M. Silva, and M.C. Cavalcanti, “Guidelines for Making Explicit the Ontological Commitment of a Legacy Database Conceptual Schema”. Submitted for publication in 2012.

[5] B. Gonçalves, “An Ontological Theory of the Eletrocardiograma with Applications”. Dissertação (Mestrado em Ciência da Computação). Universidade Federal do Espírito Santo. Espírito Santo. 2009.

[6] C. Batini, S. Ceri, S.B. Navathe, “*Conceptual Database Design: An Entity-Relationship Approach*”, Benjamin/Cummnigs, 1992.

[7] C. Welty, “OntOWLClean: Cleaning OWL ontologies with OWL”. *Proc of Formal Ontologies in Information Systems (FOIS)* (2006), 347-359, IOS Press.

[8] C.A. Heuser, “Projeto de banco de dados”, 4th edition, Sagra Luzzatto, 2001.

[9] CPRM “Cadastro dos Litotipos Estratigráficos”, Manual XIV: Manuais Téc. da CPRM. 2007.

[10] D. Calvanese, G. De Giacomo, M. Lenzerini, D. Lembo, A. Poggi, R. Rosati, “MASTRO-I: Efficient Integration of Relational Data through DL Ontologies”, *Proc. of Description Logic Workshop* (2007), 227-234.

[11] D. Yeh, L. Yuwen, W. Chu, “Extracting entity-relationship diagram from a table-based legacy database”, *Proc of The Journal of Systems and Software* **81** (2008), 764–771, Elsevier Science.

[12] E.O. Silva, J. Lisboa Filho, G. Gonçalves, "Improving Analysis Patterns in the Geographic Domain Using Ontological Meta-Properties", *Proc. of International Conference on Enterprise Information Systems (ICEIS)* **3** (2008), 256-261.
[13] E.R. Sacramento, V.M.P Vidal, J.A. Macêdo, B.F.Lóscio, F.L. Lopes, M.A. Casanova, F. Lemos, "Towards Automatic Generation of Application Ontologies". Proc of. 12th International Conference on Enterprise Information System (ICEIS) (2010), 403-406, SciTePress.
[14] F. Cerbah, "Learning highly structured semantic repositories from relational databases: the RDBtoOnto tool". In: *Proceedings of the 5th European Semantic Web Conference on the Semantic Web (ESWC)*, pages 777–781. Springer-Verlag. 2008.
[15] F. Fonseca, and J. Martin, "Learning the Differences Between Ontologies and Conceptual Schemas Through Ontology-Driven Information Systems," *Proc of Journal of the Association for Information Systems (JAIS) - Special Issue on Ontologies in the Context of IS* **8** (2007), 129–142.
[16] G. Guizzardi, "On the Representation of Quantities and their Parts in Conceptual Modeling", *Proc of 6th Int. Conf. on Formal Ontologies in Information Systems (FOIS*) (2010), 103-116, IOS Press.
[17] G. Guizzardi, "Ontological Foundations for Structural Conceptual Models", PhD. thesis , N. 05-74, Centre for Telematics and Information Technology, Univ. of Twente, Enschede, The Netherlands, 2005.
[18] G. Guizzardi, "The Problem of Transitivity of Part-Whole Relations in Conceptual Modeling Revisited", *Proc of 21st Int. Conf. on Advanced Inform. Syst. Engineering (CAISE)* **8** (2009), 94-109, Springer.
[19] G. Guizzardi, and G. Wagner, "Towards Ontological Foundations for Agent Modeling Concepts using UFO". *Proc of Agent-Oriented Information Systems (AOIS)* (2005), 110-124.
[20] G. Guizzardi, M. Lopes, F. Baião, R.A. Falbo, "On the importance of Truly Ontological Distinctions for Ontology Representation Languages: An Industrial Case Study in the Domain of Oil and Gas", *Proc of Enterprise, Business-Process and Information Systems Modeling, 10th International Workshop, (BPMDS)* (2009) *and 14th Int. Conf., (EMMSAD)* (2009), held at *(CAiSE)* (2009), 224-236, Springer.
[21] G. Guizzardi, R.A. Falbo, R.S.S. Guizzardi, "Grounding Software Domain Ontologies in the Unified Foundational Ontology (UFO): The case of the ODE Software Process Ontology", *Proc of the XI Iberoamerican Workshop on requirements engineering and software environments* (2008), 127-140.
[22] G. Wiederhold, "Mediators in the Architecture of Future Information Systems". *Proc of IEEE Computer* **25** (1992), 38-49.
[23] H. Wache, T. Vögele, U. Visser, H. Stuckenschmidt, G. Schuster, H. Neumann, S. Hübner, "Ontology-based Integration of Information - A Survey of Existing Approaches". *Proc of Workshop: Ontologies and Information Sharing, (IJCAI)* (2001), 108-117.
[24] I.M. Delgado, "A organização e atualização do léxico estratigráfico do Brasil por meio do geobank", *Proc of 44° Congresso Brasileiro de Geologia* (2008), 26-31.
[25] INDE, "Plano de Ação para Implantação da INDE - Infraestrutura Nacional de Dados Espaciais", *Proc of* http://www.inde.gov.br, 2010.
[26] J.H. Gonçalves, "Geobank e arcexibe, trabalhando conexões", *Proc of 44° Congresso Brasileiro de Geologia* (2008), 26-31.
[27] L. Lubyte, and S. Tessaris, "Automatic extraction of ontologies wrapping relational data sources. In *Proceedings of the 20th International Conference on Database and Expert Systems Applications (DEXA)*, pages 128–142. 2009.
[28] M.L.B. Villela, "Validação de Diagramas de Classe por Meio de Propriedades Ontológicas". M.Sc. Dissertation (Mestrado em Ciência da Computação). Univ. Fed. de Minas Gerais, Brazil, 2004.
[29] M.T. Özsu, P. Valduriez, "Principles of Distributed Database Systems". 2nd. ed. Prentice-Hall. 1999.
[30] N. Guarino, "Formal Ontology and Information Systems". *Proc of Formal Ontologies Information Systems*, N. Guarino (Ed.), IOS Press, 3 -15, 1998.
[31] N. Guarino, C. Welty, "A formal ontology of properties", *Proc of 12th Int. Conf. On Knowledge Engineering and Knowledge Management* (2000), 97-112, Springer Verlag.
[32] N. Guarino, C. Welty, "Evaluating Ontological Decisions with ONTOCLEAN", *Communications of the ACM* **5** (2002), 61-65.
[33] OGC, "OpenGIS Consortium Inc.", available at http://www.opengeospatial.org/, 2005.
[34] R, Raban, B. Garner, "Ontological engineering for conceptual modeling", *Proc of the 24th German/9th Austrian Conference on Artificial Intelligence*, (2001), 16-29.
[35] R. Alhajj, "Extracting the extended entity-relationship model from a legacy relation database", *Information Systems* **28** (2003), 597-618, Elsevier Science.
[36] S.P. Parker, "McGraw-Hill dictionary of scientific and technical terms", 4ª ed. New York: MacGraw-Hill Book. 1989.
[37] V.S. Silva, M.L.M. Campos, J.C.P. Silva, M.C. Cavalcanti, "An Approach for the Alignment of Biomedical Ontologies based on Foundational Ontologies", *Proc. of Journal Information and Data Management (JIDM)* **2** (2011), 557- 572.
[38] W. Kent, "Data and Reality", North Holland, 2nd ed., 2000.