What Geologists Talk About: Towards a Frequency-Based Ontological Analysis of Petroleum Domain Terms

Luan Fonseca Garcia[1], Fabricio Henrique Rodrigues[1], Alcides Gonçalves Lopes Júnior[1], Renata dos Santos Alvarenga[2], Michel Perrin[3], Mara Abel[1]

[1]Informatics Institute, Federal University of Rio Grande do Sul, Porto Alegre, Brazil

[2]Geosciences Institute, Federal University of Rio Grande do Sul, Porto Alegre, Brazil

[3]Geosiris, Paris, France

{luan.garcia, fhrodrigues, rsakuchle, agljunior, marabel}@inf.ufrgs.br

mjmperrin@gmail.com

Abstract. *One of the main challenges in building a domain ontology is to define what are the set of relevant terms to formally define and include in the conceptual model. The goal is that the selected terms provide good coverage on the required view over the domain and represent the common shared knowledge. In this work, we combine the examination of a technical thesaurus with a corpus of scientific papers of the domain to define a set of relevant geological terms as the first step towards developing a domain ontology for Petroleum Geology. To support the decision of which terms from the thesaurus we should analyze and model first, we ranked the terms on the corpus according to an adaptation of the TF-IDF statistic. After that, domain experts analyzed the top-ranked terms and selected only those that are continuants, related to Petroleum Geology. Finally, for each term chosen by the experts, we classified them as a specialization of the GeoCore Ontology or the BFO top ontology and elaborated an Aristotelian definition.*

1. Introduction

Ontologies can promote greater consistency when applied for describing and recovering data. They increase the precision of descriptions and flexibility of consultation since they can expand the queries terms to capture the user's intention. Ontologies have been used as tools for data interoperability in the petroleum industry for at least two decades. In the domain of Petroleum Geology, there are a few specialized ontologies that define very specific terms and the relationships among them within the fields of Sedimentary Geology [Lorenzatti et al. 2010, Abel et al. 2012, Carbonera 2012, Garcia et al. 2017], geological time [Cox and Richard 2005, Perrin et al. 2011, Muniz et al. 2018, Rademaker et al. 2019], Structural Geology [Babaie et al. 2006, Zhong et al. 2009], and Geological Mapping [Boyd 2016, Mantovani et al. 2020]. However, Petroleum Geology still lacks a domain ontology that defines terms used across the different geological disciplines. In the development of such ontology, ontologists and Geology experts' great effort is necessary to define what terms this general domain ontology should have, what their definitions are, and how they are related to each other.

Building a domain ontology from scratch can be a long-term work that requires several interactions and meaningful negotiation with experts and the community. A particularly hard task is defining the initial set of terms that should belong to the ontology hierarchy. In Petroleum Geology, this selection involves the definition of discipline, scale of analysis, level of detail, and immediate objective of the domain ontology, demanding long meetings and great effort form the team in charge of the task.

Several attempts from the Ontology Engineering community are trying to reduce the effort of the initial domain ontology building steps. Adopting a domain thesaurus and a frequency-based analysis is relevant in both, raising the more common technical terms and showing how useful this terminology is in state of the art, since some thesauri, although representative of consensual knowledge, can take decades to evolve. State-of-the-art works use this approach, such as [Kless et al. 2012, Cardillo et al. 2014, Kless et al. 2015, Kless et al. 2016, Kushida et al. 2017], which are not, however, in the Petroleum domain. Specifically for Petroleum, a prominent thesaurus is the GemPET [Edinger and Barker 1996]. The Geoscience, Mineral and Petroleum Thesaurus is a thesaurus owned by the Western Australian Department of Mines and Petroleum (DMP) containing 9131 English terms relative to the scope of Geosciences and Petroleum. We believe that GemPET can serve as a good basis to develop a domain ontology for Petroleum Geology. However, there are limitations when using it for this task.

The first problem is that the GemPET's objective is to provide a compilation of a shared vocabulary for indexing purposes. Thus, like any other thesaurus, it merely provides terms, but not their definitions. The second problem is that the relations contained in GemPET are very ambiguous. It contains only three types of relations (hierarchical, associative, and equivalence) that are semantically overloaded. For instance, the same type of hierarchical relations can represent relations of subtype, parthood, and instantiation.

Furthermore, although there are works such as [Kless et al. 2012, Cardillo et al. 2014, Kless et al. 2015, Kless et al. 2016, Kushida et al. 2017] that propose useful methodologies on developing an ontology from a thesaurus, these works consider that any word in the thesaurus should be part of the ontology. This decision can hardly be effective since axiomatic definitions and the terminology proper specialization are laborious work. Besides that, it leads to large prolix ontologies with reduced utility. Still, there is no guide on selecting the relevant terms for an ontology in an industrial scenario like ours.

In this work, we make a first step in the direction of developing a domain ontology for Petroleum Geology by proposing an ontological analysis of relevant geological terms selected from the GemPET thesaurus. We used an adaptation of TF-IDF statistics to rank the thesaurus terms according to their relevance in a corpus of scientific papers of the domain. We are aware that more sophisticated artificial intelligence methods are appliable, but we consider that frequency and specificity are enough to identify the relevant terms. Then, domain experts examined the top-ranked terms to identify the continuant entities relevant within Geology (i.e., those that are exclusively related to Geology, which is the scope of our ontology, instead of terms from other fields associated with petroleum such as Geophysics or Economics).

We proposed an ontological definition for each selected term in alignment with the Basic Formal Ontology and the GeoCore Ontology. This alignment is important because both ontologies aim to improve the reuse and interoperability of scientific domain ontologies. We show here an initial set of 15 terms to demonstrate the application of the methodology.

We structured the remaining of this paper in the following. In Section 2, we present the TF-IDF, BFO and GeoCore Ontology. In Section 3, we present the methodology we used to select the terms and to create their ontological definitions. In Section 4, we present the ontological definitions of the terms and their alignment with the higher-level ontologies that we use. Finally, in Section 5, we present our conclusions and next steps.

2. Related Work

In this section, we introduce the TF-IDF statistics that we used to rank GemPET terms and the GeoCore Ontology and BFO, the higher-level ontologies that encapsulate a set of ontological notions that our analysis is based.

2.1. TF-IDF

The TF-IDF (term frequency & inverse document frequency) is one of the most popular and effective term weighting strategies in information retrieval [Jones 1972], text categorization [Chen et al. 2016, Ghosh and Desarkar 2018, Zhang and Ge 2019], and text classification [Christian et al. 2016, Zhang and Chen 2020]. This approach aims to find the relevance of the term t of document d in a corpus. The TF-IDF value increases proportionally to the number of times the term t appears in document d and is offset by the number of documents in the corpus that contains this term. Equation 1 shows the traditional TF-IDF function.

$$w(t) = TF(t, d) * log(\frac{N}{DF(t)}) \tag{1}$$

where the function TF returns the frequency of the term t in document d, and the rest of the equation computes the IDF value; the function DF returns the number of documents where t occurs; N is the number of documents in the corpus.

In the traditional formulation of TF-IDF, the IDF computation uses global information on the occurrence of a given term in other documents as a strategy to limit the relevance score of the terms that occur in several documents. However, for a given term, TF-IDF computes its relevance score for a single document, i.e., if the same term occurs in different documents, each occurrence may have different relevance scores. Thus, the TF-IDF cannot find a global relevance score for a term in a corpus.

Another disadvantage is that the TF-IDF calculates the relevance score for terms with a single word. Then, if a term is a compound word, this word has no associated relevance score. In this situation, for each word that composes a compound word, it is computed a relevance score. Furthermore, when two terms are synonymous, each of them has independent relevance scores.

2.2. Basic Formal Ontology

The Basic Formal Ontology (BFO) is a small top-level ontology developed for supporting data integration of scientific research [Arp et al. 2015]. According to the authors, the goal of an ontology focused on describing a scientific domain is *"to encapsulate the knowledge of the world that is associated with the general terms used by scientists"* in this domain.

It is composed of domain-independent terms, which correspond to top terms shared between many domain ontologies. BFO thus provides a top-level structure that allows the integration of the information compiled by various domain ontologies in a common framework for categorization and reasoning.

We chose BFO as the top-level ontology for this work for many reasons. BFO bases its philosophic background on realism and aims in dealing with real entities of the material world, which is convenient for a Geology ontology. BFO supports a kind of incremental development where we define new entities by specializing upper classes without worrying about a complete disjoint partition of the domain instances. This vision allows us to build "networks" of related domain ontologies that can reuse classes keeping a certain degree of modularity. These characteristics align with proper documentation that makes the industrial community advises BFO as the best top ontology for industrial applications.

2.3. GeoCore Ontology

GeoCore Ontology is a core ontology for Geology that extends the classes existent in the BFO top-ontology to define a set of upper-level entities that represent the most general and basic naturally occurring entities in the Geology domain [Garcia et al. 2020]. Its goal is to serve as a common ground for developing new ontologies by specialization or integrating existing domain ontologies in the geological domain.

In the heart of GeoCore lies the distinction between geological objects and the earth material that constitutes them. They are both independent continuants, but with distinct nature regarding their unity criteria. Geological Objects are objects (defined in BFO as a material entity that manifests causal unity) that are naturally occurring and are constituted by some Earth Material. On the other hand, an Earth Material is a material entity that is a natural amount of matter, that is solid, fluid, or unconsolidated, and that does not manifest causal unity. In this view, GeoCore separates a sedimentary layer made of sandstone into two distinct entities, the geological object (the layer) and the earth material that constitutes it (the sandstone rock).

GeoCore defines two specific classes of Earth Materials: Earth Fluids - water, oil, gas, or a mixture of that fluids-, and Rocks - solid consolidated Earth Materials made of mineral matter or biological origin. However, there is no claim that these are the only subtypes of Earth Materials that exist.

Another important class defined in GeoCore is the Geological Structure, a Generically Dependent Continuant that is the pattern of the spatial arrangement of the internal parts of a Geological Object.

3. Methodology

This work's approach consists of using GemPET thesaurus terms as a list of candidate terms to be ontologically defined. However, beyond the simple terms, we are interested

Tabela 1. Number of selected papers per journal.

Journal	Year 2019	Year 2020	Total
AAPG Bulletin	113	28	141
Brazilian Journal of Petroleum and Gas	28	0	28
Marine Petroleum Geology	117	0	117
PETROLEUM	49	0	49
Petroleum Research	29	6	35
Petroleum Science	107	19	126
Total			496

in the entities to which the terms refer in reality. Thus, using the equivalence relation that the thesaurus provides, we grouped the 9131 terms into 6338 synsets (i.e., a set of one or more synonyms terms that are interchangeable in some context). We consider each synset as referring to a single entity in reality, and this was the ”material”on which we worked.

Our methodology comprises three main phases: synset ranking, synset selection, and ontological analysis of the selected synsets. The first phase (section 3.1) involves the data preparation and ranking methods that we used to automatically decide which synsets should be prioritized in our ontological analysis - avoiding the time-consuming task of manually assessing the relevance of each of the 6338 synsets. In the second phase (3.2), after the synsets were ranked, two experts of the Geology domain examined the most top-ranked terms and selected those related to Geology. Finally, in the third phase (section 3.3), we analyzed the selected synsets and proposed an ontological definition for each one of them.

The experts show different professional profiles: one has a solid scientific theoretical background, while the other has vast experience in industrial projects in Geology. Both have some basis in Ontology Engineering and were working in ontology development in the last years.

3.1. Synset Ranking

We ranked the synsets of the thesaurus based on the statistical relevance of their equivalent terms in a corpus of scientific papers from the Petroleum domain. The underlying idea is that the synsets whose terms are the most relevant in the corpus refer to relevant entities of the field.

The corpus is composed of 496 articles selected from 6 high-impact scientific journals published in the last 18 months that cover the main geological subjects on petroleum exploration (as shown in table 1). The intention was to have a corpus that is both recent and scientific relevant for the ontology domain.

We started the ranking task by removing the non-alphanumeric characters and applying a process of stemming for every synset term. Stemming consists of reducing all inflected or derived words to their stem. Our objective with this was to improve the string matching during the first phase of our methodology. Then, we extracted the textual information of the papers corpus and applied the same textual processing we used to terms of the thesaurus.

After that, we applied an adapted version of the TF-IDF statistic in which we

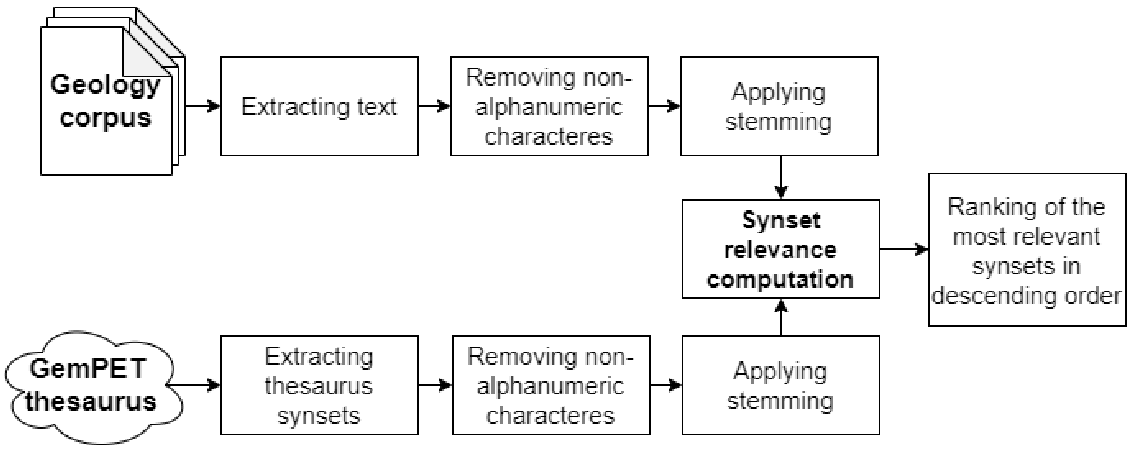

Figura 1. Methodology for ranking the synsets of the GemPET thesaurus.

named Synset Global Relevance. In the SGR approach (equation 2), for a given synset s, we return the sum of the synset frequencies SF in every document d from the set of documents N offset by the number of documents DF in which it occurs. The SF function is the sum of the frequencies of all synset s synonyms terms.

$$SGR(s) = log\left(\frac{N}{DF(s)}\right) * \sum_{j=1}^{N} SF(s, d_j) \tag{2}$$

Finally, we ranked synsets by its SGR in descending order. Thus, the best-ranked synsets are those comprising the terms considered to be the most relevant according to the corpus we used. Figure 1 presents all the steps used to obtain the ranking of the most relevant GemPET synsets in the corpus.

3.2. Synset Selection

After the ranking phase, we selected the 1% top-ranked synsets (i.e., 63 of the 6338, as presented in figure 2). We then chose a preferred term for each of the selected synsets and submitted them to the analysis of two experts from the geological domain.

The experts made this analysis in two steps. First, each expert decided about the inclusion of each term in the ontological analysis phase based on two questions:

1. Is it a term from the Petroleum Geology domain?
2. Does it refer to a continuant (i.e., an object or a quality to describe an object) rather than to an occurrent (i.e., an event, a process)?

The first question guarantees that the selected term is within the domain's defined scope, avoiding frequent words referring to economic or engineering aspects of petroleum exploration. The second question was due to our methodological approach of first defining the continuants of the domain and then defining occurrents according to their continuant participants. We included in the ontological analysis phase 15 terms (figure 2) for which experts agreed in answering yes to both questions (about 25% of the 63 initially considered terms).

TERM	SCORE	Included	
		E1	E2
Fractures	0.8114	x	x
Shale	0.5884	x	x
Faults	0.5208	x	x
Faulting	0.5177		
Basins	0.5039	x	x
Emulsions	0.4423		
Seismicity	0.4303		
Crude oil	0.4252	x	x
Shear properties	0.4129		
Clays	0.4068		x
Surfactants	0.3773		
Oil recovery	0.3673		
Catalysts	0.3540		
Stress	0.3491		
Tectonics	0.3457		
Source beds	0.3413	x	x
Recovery	0.3395		
Facies	0.3384	x	x
Brines	0.3307		
Polymerization	0.3270		
Polymers	0.3270		

TERM	SCORE	Included	
		E1	E2
Prices	0.3235		
Asphaltene	0.3162		
Absorption	0.3107		
Porosity	0.3065	x	x
Hydration	0.3065		
Viscosity	0.3038		
Pressure	0.2981	x	
Well cementing	0.2956		
Hydraulic fracturing	0.2919		x
Permeability	0.2886	x	x
Cementation process	0.2837		
Cements	0.2829	x	x
Fuels	0.2795		
TOC	0.2721		x
Pipelines	0.2720		
Deposition	0.2715	x	
Diffusion	0.2702		
Sandstone	0.2677	x	x
Floods	0.2677		
Mineralization	0.2657		
Minerals	0.2657	x	x

TERM	SCORE	Included	
		E1	E2
Biomarker hydrocarbons	0.2619		x
Organizations	0.2545		
Dolomitization	0.2476		
Salinity	0.2424		
Salinization	0.2424		
Acidity	0.2313		
Acids	0.2313		
Logging	0.2289		
Logs	0.2289		
Slurries	0.2273		
Temperature	0.2268	x	
Sedimentation	0.2250		
Sediments	0.2218	x	x
Core drilling	0.2217		
Organic material	0.2215	x	x
Sampling	0.2209		
Sands	0.2196	x	x
Petroleum geology	0.2171		
Maturation	0.2169		
Hydrodynamics	0.2168		
Water flooding	0.2162		

Figura 2. Terms included in the ontological analysis.

In the second step, both experts selected, discussed, and improved each included term's definitions based on the Schlumberger Oilfield Glossary[1], Encyclopedia Britannica[2], Glossary of Geology [Jackson 2005], and Wikipedia.

3.3. Ontological Analysis

In this work, we conducted an ontological analysis to identify and model the entities of reality to which the included terms refer. We based this analysis on the definitions provided by the experts. For each term, we merged the definitions of both experts, considering their common points with ontological relevance (e.g., related to essential properties of the entity rather than accidental properties) and the inconsistencies on the public glossaries. Then, we analyzed each of the merged definitions in the light of GeoCore Ontology and BFO to identify which class of these ontologies the definition specializes in.

After identifying what class would be specialized, we elaborated an Aristotelian definition in natural language for each of the included terms. An Aristotelian definition has the form **A =def is a B that C**, where **A** is the term we are defining, **B** is a class of GeoCore or BFO, and **C** is the set of properties that makes **B** a specialization of **A**.

During this process, we identified 2 additional terms (Sedimentary Rock and Unconsolidated Earth Material) that were not present in the initial list and that required a definition (e.g., terms referring to entities that directly subsume entities referred by the included term). Thus, we also added those to the list and asked for definitions from the experts. For each additional term, we created its Aristotelian definition, which we submitted to the experts for possible adjustments. We present the result of the ontological analysis in the following section.

4. Definitions

Section 4.1 contains the ontological definitions of the terms selected by the experts following the methodology we presented in the previous section. In addition to the 15 terms

[1]https://www.glossary.oilfield.slb.com/
[2]https://www.britannica.com/

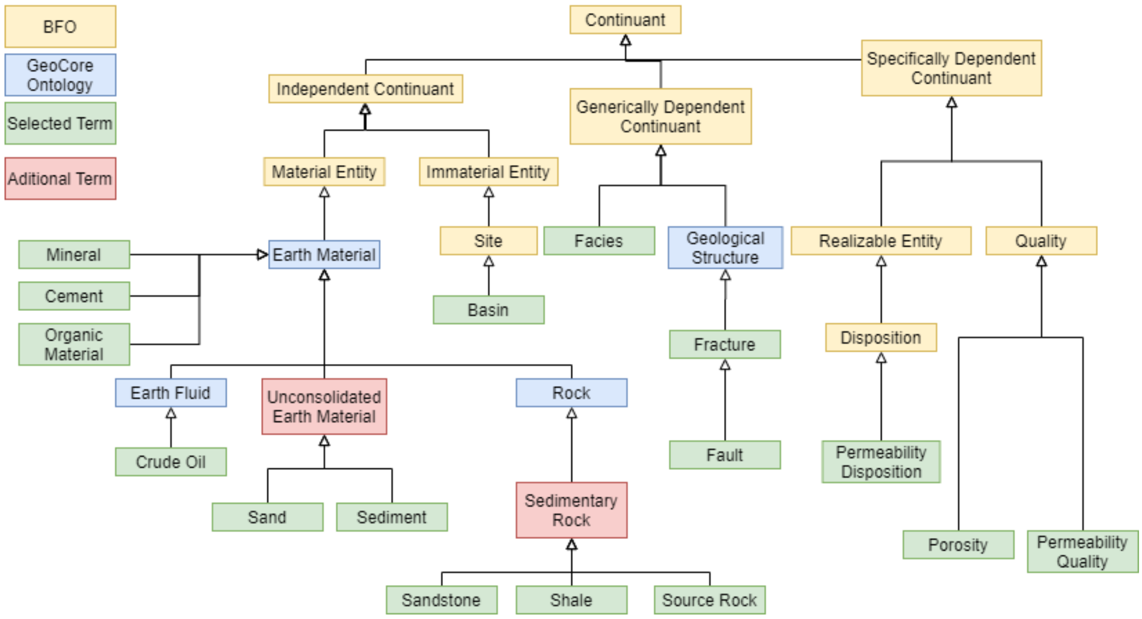

Figura 3. Relations of subsumption between the terms defined and the GeoCore and BFO ontologies.

chosen initially, we needed to define two more terms to create the definitions properly and one additional term for disambiguating the term permeability. In section 4.2, we discuss some aspects of the definitions that we think worth noting. Figure 3 presents all the terms with their subsumption relations either to classes from GeoCore or BFO.

4.1. Definitions

- **Basin** =Def. is a BFO:Site that is a depression placed on the Earth Crust generated by subsidence caused by tectonic activity.
- **Cement** =Def. is an GC:Earth Material that is constituted by an aggregate of crystals, that is generated by a chemical precipitation process, and that has the role of binding the grains of a sedimentary rock.
- **Crude Oil** =def. is an GC:Earth Fluid that is a naturally occurring liquid mixture of hydrocarbon compounds.
- **Facies** =def. is a BFO:Generically Dependent Continuant that is concretized by a set of BFO:Qualities of the GC:Rock in which it is concretized.
- **Fracture** =def is a GC:Geological Structure that is concretized by a topological mechanical discontinuity in one or several connected GC:Geological Objects.
- **Fault** =def. is a Fracture that is approximately planar and is concretized by a displacement between two GC:Geological Objects or two BFO:Object Fiat Parts of a GC:Geological Object.
- **Mineral** =Def. is an GC:Earth Material that is a naturally occurring, inorganic, solid, homogeneous chemical compound with crystalline structure.
- **Organic Material** =Def. is an GC:Earth Material that is composed of biological carbon-based compounds.
- $\textbf{Permeability}_{\textbf{disposition}}$=def is a BFO:Disposition of a porous GC:Earth Material to allow fluids to pass through it.

- **Permeability$_{\text{quality}}$**=def is a BFO:Quality of a porous GC:Earth Material that is the interconnectivity between its void spaces.
- **Porosity** =def is a BFO:Quality of GC:Rock or Unconsolidated Earth Material that is the ratio of the volume of the empty spaces and the total volume of the material.
- **Sand** =Def. is an Unconsolidated Earth Material that is constituted by an aggregate of detrital grains finer than gravel and coarser than silt[3].
- **Sedimentary Rock** =Def. is a GC:Rock that is generated by the lithification of sediments or by mineral precipitation from a chemical solution in normal surface temperature.
- **Sandstone** =Def. is a Sedimentary Rock that is constituted by a consolidated aggregate of mainly sand-sized siliciclastic grains of rock or mineral particles.
- **Sediment** =Def. is an Unconsolidated Earth Material that bears the role of being deposited at the surface of the Earth.
- **Shale** =Def. is a Sedimentary Rock constituted by a consolidated aggregate of laminas composed of mainly siliciclastic silt-sized grains.
- **Unconsolidated Earth Material** =Def. is an GC:Earth Material that is constituted by an aggregate of solid particles that is not consolidated into a rock itself.
- **Source Rock** =Def. is a Sedimentary Rock that is composed of organic material and that bears the role of being the host of the transformation of this organic material into hydrocarbons.

4.2. Discussion

During the ontological analysis we identified some aspects of the terms that are worth noting.

Firstly, the analysis allowed us to identify two distinct entities hidden due to semantic overload in the term permeability. Permeability may refer to the degree of interconnection between the void spaces in a rock or to the disposition based on such interconnectivity, which makes the rock able to allow the passage of fluids. When we consider them distinct, we highlight a point of integration to define occurrents based on the realization of this disposition.

Another contribution of the ontological analysis was evidencing the distinction between the material entities of the domain (e.g., rock, sand) and the roles they may play (e.g., source rock, sediment, cement). This distinction is important because we can differentiate what an entity essentially is from the characteristics that it may acquire during its existence. We consider it another point of integration to define occurrents as the state transitions of such entities.

Lastly, an interesting fact to point out is that all terms that we defined are specializations of some class of GeoCore Ontology or directly depend on some class of it (e.g., in the case of qualities, dispositions, and roles).

5. Concluding Remarks

The first problem in building domain ontologies from scratch is identifying the relevant initial terminology that should be the focus of modeling and will eventually define the

[3]Gravels are particles with size ranging from 2 mm to 63mm and silts are particles with size ranging from 0.002 mm to 0.063mm.

main subsumption structure of the ontology. In this work, we defined a set of initial relevant geological terms towards developing a domain ontology for Petroleum Geology to support retrieving information over a repository of geosciences documents. To support the decision of which terms from the thesaurus we should analyze first, we ranked them according to an adaptation of the TF-IDF statistic over a corpus of scientific papers of the domain. After that, domain experts analyzed the top-ranked terms and selected only those that are continuants and that are related to Petroleum Geology. Finally, for each term chosen by the experts, we aligned it with the GeoCore Ontology and the BFO, and we elaborated on an Aristotelian definition.

Currently, there are a few specialized well-founded ontologies in the geological domain and the field still lacks a more general domain ontology that defines domain terms used across the different geological disciplines. Developing such an ontology from scratch is a resource-consuming task, requiring the dedication of domain experts who usually have a restricted time availability. On the other hand, simply using a thesaurus as the single source for ontology development is not realistic due to the absence of definitions of the terms, and the ambiguity of the relations contained on it, still requiring the support of experts. Thus, with this work we aimed to show that ranking the thesaurus terms according to their relevance within a corpus of the domain is a helpful approach to optimize the available resources (e.g., the time of experts), initially selecting a small set of terms to analyze from thesauri that have a considerable number of terms.

One may say that the statistical relevance of the terms does not necessarily match with the relevance given by the experts of the domain. Still, it seems impractical to work with more than thousands of terms otherwise. Furthermore, the set of included terms covers important entities in the field (e.g., sedimentary rock, sediment, petroleum), their functions (e.g., source rock), and characteristics (e.g., porosity, permeability). Additional evidence of the practical relevance of the terms we selected is their presence in manually-constructed ontologies of the domain (e.g., fault and fracture in [Zhong et al. 2009, Boyd 2016], porosity and permeability in [Lorenzatti et al. 2010, Garcia et al. 2017], and facies in [Carbonera 2012]). Besides that, the broad distribution of classes corresponding to the defined terms in the classes of GeoCore and BFO (including, for example, Earth Material, Qualities, Immaterial Entities) indicates that it is not the case of covering just a tiny part of the domain.

Our work presents some limitations. Currently, it just presents an ontological analysis of a set of terms that are arguably relevant in the domain that may be used as a starting point to build a domain ontology, but that still cannot be considered a proper ontology itself. Moreover, we did not contemplate terms referring to occurrents. Therefore, as future work, we plan to further apply the proposed methodology to enlarge the list of selected terms, including those referring to occurrents, towards the development of an ontology for the domain.

Acknowledgments

We thank the Brazilian Agencies CAPES and CNPQ for the support of this project. This work is part of the Digital Geology project, a research cooperation with Petrobras company.

Referências

Abel, M., Lorenzatti, A., Ros, L. F. D., da Silva, O. P., Bernardes, A., Goldberg, K., and Scherer, C. (2012). Lithologic logs in the tablet through ontology-based facies description. *AAPG Annual Convention and Exhibition*.

Arp, R., Smith, B., and Spear, A. D. (2015). *Building ontologies with basic formal ontology*. Mit Press.

Babaie, H. A., Oldow, J. S., Babaei, A., Lallemant, H. G. A., Watkinson, A. J., and Sinha, A. (2006). Designing a modular architecture for the structural geology ontology. *Sepcial Papers - Geological Society of America*, 397:269.

Boyd, T. (2016). Ontology of geological mapping. Master's thesis, College of Arts and Sciences, Georgia State University.

Carbonera, J. L. (2012). Raciocinio sobre conhecimento visual: Um estudo em estratigrafia sedimentar. Master's thesis, Universidade Federal do Rio Grande do Sul (UFRGS).

Cardillo, E., Folino, A., Trunfio, R., and Guarasci, R. (2014). Towards the reuse of standardized thesauri into ontologies. In *WOP*, pages 26–37.

Chen, K., Zhang, Z., Long, J., and Zhang, H. (2016). Turning from tf-idf to tf-igm for term weighting in text classification. *Expert Systems with Applications*, 66:245 – 260.

Christian, H., Agus, M. P., and Suhartono, D. (2016). Single document automatic text summarization using term frequency-inverse document frequency (tf-idf). *ComTech: Computer, Mathematics and Engineering Applications*, 7(4):285–294.

Cox, S. J. and Richard, S. M. (2005). A formal model for the geologic time scale and global stratotype section and point, compatible with geospatial information transfer standards. *Geosphere*, 1(3):119–137.

Edinger, J. and Barker, T. A. (1996). *GeMPeT: the geoscience, minerals and petroleum thesaurus*. Charted Information Services.

Garcia, L. F., Abel, M., Perrin, M., and dos Santos Alvarenga, R. (2020). The geocore ontology: A core ontology for general use in geology. *Computers & Geosciences*, 135:104387.

Garcia, L. F., Graciolli, V., De Ros, L. F., and Abel, M. (2017). A conceptual framework for rock data integration in reservoir models based on ontologies. *International Journal of Monitoring and Surveillance Technologies Research (IJMSTR)*, 5(1):71–82.

Ghosh, S. and Desarkar, M. S. (2018). Class specific tf-idf boosting for short-text classification: Application to short-texts generated during disasters. In *Companion Proceedings of the The Web Conference 2018*, page 1629 to 1637, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Jackson, J. A. (2005). Glossary of geology. *Glossary of Geology, by JA Jackson. 2005 Approx. 900 p. 5th revised and enlarged ed. ISBN 3-540-27951-2. Berlin: Springer, 2005.*, page 5th.

Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*.

Kless, D., Jansen, L., Lindenthal, J., and Wiebensohn, J. (2012). A method for re-engineering a thesaurus into an ontology. In *FOIS*, pages 133–146.

Kless, D., Jansen, L., and Milton, S. (2016). A content-focused method for re-engineering thesauri into semantically adequate ontologies using owl. *Semantic Web*, 7(5):543–576.

Kless, D., Milton, S., Kazmierczak, E., and Lindenthal, J. (2015). Thesaurus and ontology structure: Formal and pragmatic differences and similarities. *Journal of the Association for information science and technology*, 66(7):1348–1366.

Kushida, T., Kozaki, K., Tateisi, Y., Watanabe, K., Masuda, T., Matsumura, K., Kawamura, T., and Takagi, T. (2017). Efficient construction of a new ontology for life sciences by sub-classifying related terms in the japan science, technology agency thesaurus. In *ICBO*.

Lorenzatti, A., Abel, M., Fiorini, S. R., Bernardes, A. K., and dos Santos Scherer, C. M. (2010). Ontological primitives for visual knowledge. In *Proceedings of the 20th Brazilian conference on Advances in artificial intelligence (2010)*, volume 6404 of *Lectures Notes in Artificial Intelligence*, pages 1–10, Sao Bernardo do Campo. Springer Berlin / Heidelberg.

Mantovani, A., Lombardo, V., and Piana, F. (2020). Ontology-driven representation of knowledge for geological maps. *Computers & Geosciences*, page 104446.

Muniz, H., Chalub, F., Rademaker, A., and de Paiva, V. (2018). Extending wordnet to geological times. In *Proceedings of the 9th Global WordNet Conference (GWC 2018)*, page 146.

Perrin, M., Mastella, L. S., Morel, O., and Lorenzatti, A. (2011). Geological time formalization: an improved formal model for describing time successions and their correlation. *Earth Science Informatics*, 4(2):81–96.

Rademaker, A., Tessarollo, A., Muniz, H., and Pease, A. (2019). Extending sumo to geological times. *ONTOBRAS*.

Zhang, T. and Chen, C. (2020). Research on automatic text summarization method based on tf-idf. In Xhafa, F., Patnaik, S., and Tavana, M., editors, *Advances in Intelligent Systems and Interactive Applications*, pages 206–212, Cham.

Zhang, T. and Ge, S. S. (2019). An improved tf-idf algorithm based on class discriminative strength for text categorization on desensitized data. In *Proceedings of the 2019 3rd International Conference on Innovation in Artificial Intelligence*, ICIAI 2019, pages 39 – 44, New York, NY, USA. Association for Computing Machinery.

Zhong, J., Aydina, A., and McGuinness, D. L. (2009). Ontology of fractures. *Journal of Structural Geology*, 31(3):251–259.