Fundação Getulio Vargas
Escola de Matemática Aplicada

Alexandre Tessarollo

Expanding the Open Wordnets for English and Portuguese to Geology Domain: Inclusion of Lythology and Geological Time Concepts

Rio de Janeiro
2020

Alexandre Tessarollo

Expanding the Open Wordnets for English and Portuguese to Geology Domain: Inclusion of Lythology and Geological Time Concepts

Dissertação submetida à Escola de Matemática Aplicada como requisito parcial para a obtenção do grau de Mestre em Modelagem Matemática.

Orientador: Alexandre Rademaker

Rio de Janeiro
2020

Dados Internacionais de Catalogação na Publicação (CIP)
Ficha catalográfica elaborada pelo Sistema de Bibliotecas/FGV

Tessarollo, Alexandre
Expanding the open Wordnets for english and portuguese to geology domain : inclusion of lythology and geological time concepts / Alexandre Tessarollo. – 2020.
94 f.

Dissertação (mestrado) -Fundação Getulio Vargas, Escola de Matemática Aplicada.
Orientador: Alexandre Rademaker.
Inclui bibliografia.

1. Processamento da linguagem natural (Computação). 2. Petróleo e gás. 3. WordNet. I. Rademaker, Alexandre. II. Fundação Getulio Vargas. Escola de Matemática Aplicada. III. Título.

CDD – 006.35

Elaborada por Márcia Nunes Bacha – CRB-7/4403

ALEXANDRE TESSAROLLO

"EXPANDING THE OPEN WORDNETS FOR ENGLISH AND PORTUGUESE TO GEOLOGY DOMAIN: INCLUSION OF LYTHOLOGY AND GEOLOGICAL TIME CONCEPTS''.

Dissertação apresentado(a) ao Curso de Mestrado em Modelagem Matemática do(a) Escola de Matemática Aplicada para obtenção do grau de Mestre em Modelagem Matemática.

Data da defesa: 30/09/2020

ASSINATURA DOS MEMBROS DA BANCA EXAMINADORA

Presidente da Comissão Examinadora: Profº **Alexandre Rademaker**

Alexandre Rademaker
Orientador

Mara Abel
Membro

Francis Charles Bond
Membro

Nos termos da Lei nº 13.979 de 06/02/20 - DOU nº 27 de 07/02/20 e Portaria MEC nº 544 de 16/06/20 - DOU nº 114 de 17/06/20 que dispõem sobre a suspensão temporária das atividades acadêmicas presenciais e a utilização de recursos tecnológicos face ao COVID-19, as apresentações das defesas de Tese e Dissertação, de forma excepcional, serão realizadas de forma remota e síncrona, incluindo-se nessa modalidade membros da banca e discente.

César Leopoldo Camacho Manco
Diretor

Antonio de Araujo Freitas Junior
Pró-Reitor de Ensino, Pesquisa e Pós-Graduação FGV
Antonio Freitas, PhD
Pró-Reitor de Ensino, Pesquisa e Pós-Graduação
Fundação Getúlio Vargas

Instrução Normativa nº 01/19, de 09/07/19 - Pró-Reitoria FGV
Em caso de participação de Membro(s) da Banca Examinadora de forma não-presencial*, o Presidente da Comissão Examinadora assinará o documento como representante legal, delegado por esta I.N.
*Skype, Videoconferência, Apps de vídeo etc

If I have seen further, it is by standing on the shoulders of giants.

Isaac Newton [64]

Acknowledgements

First and foremost I thank God for the blessings and guidance.

I thank Petrobras for the opportunity and to the people who helped make this possible. Special thanks to my managers and coworkers, a list to extensive to enumerate here.

I thank my advisor Alexandre Rademaker for introducing me to the fabulous realm of Natural Language Processing and for all the knowledge bestowed upon me - only a fraction of it made to this dissertation, but all of it made me a better researcher and a better professional. Thanks for investing so much of your time in me.

I thank the members of my thesis committee, Francis Bond and Mara Abel. Your feedback took this material to the next level.

I thank all professors and students at FGV/EMAp for the lessons learned and all the fruitful discussions. Special thanks to the colleagues in our NLP group: Alessandra Cid, Bruno Cuconato, and Henrique Muniz.

I thank the FGV/EMAp staff, particularly Cirlei de Oliveira, Elisângela Santana, and Mônica Souza for the continuous assistance.

I thank my parents and relatives for the education and for providing the building blocks during my formation years.

Finally, I thank my wife for her unwavering support throughout this entire process, and to my beautiful boy for his patience. This work is for you.

Abstract

Human knowledge has been stored, transferred and built upon by written means. The human ability to tap into this source is by far the main reason why we've been able to advance our collective understanding. Over a quarter century ago, our technologies for collecting, storing, and disseminating vast amounts of information had gotten ahead of our technologies for collating and analyzing it. Natural Language Processing (NLP) tackles this issue. The everyday life already benefits from NLP, with applications ranging from spam filtering to (limited) support chatbots and artificial intelligence assistants interacting through voice commands. When it comes to technical language, however, NLP has some shortcomings. This is particularly true for the Oil&Gas domain, where information is the most precious resource, one that supports decisions worth billions of dollars. Even though there are numerous reports, papers, documents and alike, such knowledge remains untapped due to NLP domain limitations. It is our hypothesis that expanding a lexical resource, namely the WordNet, will have a scalable effect particularly on Word Sense Disambiguation (WSD) and on the overall NLP for Oil&Gas domain documents. To verify this we extended the WordNet with 377 new concepts (synsets), 558 new lexical forms (words) and 948 new relations (pointers) between such word and/or synsets. Such extension is focused on two of the most common references mentioned in Oil&Gas documents: Geological Time and Lithology (branch of geology devoted to rocks). We perform such extension both "vertically" from the original Princeton WordNet in English into the Open WordNet for English (OWN-EN) and "horizontally" by translating and adapting such effort to the Open WordNet for Portuguese (OWN-PT). We then compare the outputs of the WSD algorithm UKB before and after such extension. Both WordNet extensions (English and Portuguese) are available as online open-source initiatives.

Keywords: WordNet, domain extension, rock, lithology, geological time, geology, NLP, SUMO, oil and gas, petroleum, geoscience, Portuguese, English, International Chronostratigraphic Chart

Contents

Chapter 1

Introduction

Throughout millennia, human knowledge has been stored, transferred and built upon by written means. The human ability to tap into this source is by far the main reason why we've been able to advance our collective understanding. Everyday more people and more systems produce more content, which in turn result into more people and more systems producing even more content - from replies and extensions to the original ones to content on how to handle so much content. Over two and a half decades ago [56] already stated that our technologies for collecting, storing, and disseminating vast amounts of information had gotten ahead of our technologies for collating and analyzing it.

Aiding humanity in this ever-surmounting task we have computers, particularly a branch of computer science named computational linguistics, from which arises Natural Language Processing (NLP). NLP also drains from philosophy, linguistics and other fields of study that seek to understand how humans interact relying on natural language (whether written, spoken or else) and how could a computer could approach human performance levels on tasks involving natural language [2].

On day to day tasks we are already used with NLP systems supporting us. Our e-mail alone filters out most of the spam, checks for events we got enrollment confirmation to update our calendar and can even suggest short phrases when we reply e-mails. Automatic translators are already supporting travelers with real time translation, and Google search recognizes entities and concepts within our inputs, returning Wikipedia links and infoboxes.

In Oil&Gas domain information is the most precious resource. It is an industry that bases decisions worth billions of dollars on great amounts of

uncertainty due to scarce information. However, there is a huge source of untapped knowledge hidden within different set of documents, reports, scientific articles, business intelligence articles and so on. Most of the processing of this fundamental data is done by human professionals actually reading it rather than by a computational system. Because this data mountain is growing exponentially, managing it and finding relevant content quickly has become one of companies and professionals most critical challenges [6, 7, 32, 69, 86]. Hence the need and the opportunity to contribute to NLP applied to Oil&Gas domain.

1.1 NLP for Oil&Gas

Oil&Gas, however, is a quite extensive domain, with many sub-domains such as drilling, completion, seismic, sub-sea engineering, lithology, petrography, appraisal, production, refinement, and so on. Some of this areas correlates with others - for instance, seismic is an appraisal technique, but it is also used for production management later in a field's life, while also standing for an universe of concepts in itself: logging while drilling (LWD), crosswell seismic, borehole seismic, permanent reservoir monitoring system, depth-velocity conversion, among others.

This poses several challenges to Oil&Gas domain NLP modelling. Trying to embrace such a huge knowledge area resorting to domain specific literature would be a herculean task by all means. Even building a corpus would be a challenge: how to represent, let alone balance all sub-areas? On the other hand, different sub-areas sometimes use the same words with quite different meanings. The term "poor gas", for instance, might refer to a volume of natural gas from which most liquid fractions have already been extracted or to a volume of natural gas that hasn't gone through such extraction but also presents high contaminants ratio. Even though there are many NLP resources and techniques that could handle some of such matters, it would just be more encumbrances to an already hard enough task.

So, as a start, we chose to tackle with petroleum geology. On the one hand is a good choice as any, on the other hand it is the sub-domain where the whole Oil&Gas starts. But petroleum geology is also a rather extensive domain. Instead of dwelling on whether to focus on sedimentology, seismology or some other subdomain, we chose to focus on two of the most common references mentioned in Oil&Gas documents: Geological Time and Lithol-

ogy. By targeting such transverse concepts of the Oil&Gas domain, we aim to find some consensus or at least some reference authority on said concepts.

Lithology comprises loose- and bed-rock, classified according to their modal composition or their grain size, respectively[1]. Other definitions states it "is the description of rock composition (what it is made of) and texture"[2] or it is "the scientific study and description of rocks, especially at the macroscopic level, in terms of their color, texture, and composition" [30].

Geological Time is organized in the geologic timescale. It is used by geologists, paleontologists and other geoscientists to describe the timing and relationships of events in Earth's history. The table of geologic time spans set forth by the International Commission on Stratigraphy (ICS), a subcommittee of the International Union of Geological Sciences, is described in http://www.stratigraphy.org. The geologic timescale is organized in a hierarchical fashion. Eons (or aeons) are divided into eras. Eras contain periods that contain epochs, and finally epochs contain ages. The first three eons (Hadean, Archean, Proterozoic) are collectively referred as the Precambrian super-eon. The most recent eon, the Phanerozoic is subdivided into several periods.

The International Commission on Stratigraphy publishes regularly the International Chronostratigraphic Chart (ICC)[3] as the current standard of the organization of the geologic timescale of the Earth. In the current version, the chart contains 178 names of geological periods. One can read about the development of the chart in [25].

Given the choice to contribute towards Geological Time and Lithology there is the issue on where in a natural language processing "pipeline" to intervene.

1.2 NLP as a pipeline

Natural Language Processing consist of several tasks, which may be structured as a pipeline. Hence, a theoretical NLP pipeline comprehends capturing the data in text format, splitting it into sentences and words, identifying

[1] https://thesaurus.geolba.ac.at

[2] urlhttps://earthquake.usgs.gov/learn/glossary/?term=lithology, last accessed September 30 2020

[3] It was previous called International Stratigraphic Chart (ISC). It can be found at http://www.stratigraphy.org/index.php/ics-chart-timescale.

tokens, lemmas and part-of-speech, named entities, multi-word expressions and senses, all of these just to understand the information within a message. An actual pipeline may include some or all of these steps in a myriad of orders and combinations. There are several different possibilities, the example set here is for didactic purpose alone. The reader interested in parts or the whole of such pipelines should refer to [48].

The capturing into text format may include from speech to text (STT) to optical character recognition (OCR) over scanned documents. STT is used when processing audio inputs, such as a person talking to its cell phone personal assistant. One relevant concern in STT is phonetics: for instance, how to differentiate "hi" from "high" or "cell" from "sell". OCR does not face such issues, but on the other hand is concerned with differentiating O (as in Oscar), D (as in Delta) and 0 (the number). Other fields handling the input include processing PDF files (a simple copy/paste result in broken text, with random symbols and paragraphs within the text), web-crawling techniques and even text-crawling techniques - how to differentiate blocks of text from tables, figures, quotes, titles and other types of noise. In this simple explanation, let's consider they all end outputting a text file that contains only the text blocks from the original message.

The next step is splitting the text in sentences and then into words. Contrary to common sense, it is not enough to just identify periods, question marks and alike: one will face colons, commas and other punctuation like in the examples "Mr. Smith and John J. Holt we willing to pay 10.95 per share" and "J. J. Abrams is a famous writer". Other issues would be references (such as the bibliographical in this text), chemical formulas, etc. These are some relevant points for English, but some other languages may present even more issues on sentence splitting.

Once the text is splitted into sentences, there is still the need to split it into words, also known as tokenization. Hyphenation and syllables divided at line-breaks are issues, as are formulas (chemical, mathematical, etc). Other related issues are Named Entities Recognition (NER) and Multi Word Expressions (MWE). NER concerns with identifying named entities such as "John" and "Austria". Among many issues, it has to deal with common nouns which are also proper nouns such as major oil companies "Shell" and "Total". Other issue is when names are used as regular words, such as "just google it" - notice this last example turned a proper noun into a verb. MWE focus on identifying which word should be handled in a grouped form like "United States of America" and "living room". As the USA example points

out, MWE may have overlapping with NER.

Other relevant concern is the lemmatization of words, where different forms of a word are associated with their dictionary form, like seeing and saw to see. An alternative to lemmatization is stemming, which reduces inflected and derived words to their root form. Depending on the algorithm, "arguing" and "argues" are reduced to "argu".

Other steps are Part of Speech recognition (PoS) and semantic role labeling (SRL), which aims to recognize grammatical aspects such as nouns, verbs, adjectives (for PoS) and their relations (for SRL). For instance, a noun can be the subject or the direct object of a given verb. Syntactic analysis, or parsing, goes deeper, by building and choosing possible parse trees, identifying syntactical relations between the words of a given sentence.

Finally, there is Word Sense Disambiguation (WSD), which aims to associate words with their proper senses in a given reference resource. For instance, the noun *chair* may be linked to the sense *a seat for one person, with a support for the back* or to *the officer who presides at the meetings of an organization* or to one of the many other senses chair has. A major online resource commonly used for WSD is Princeton WordNet[34][4]. Chapter 2 describes Princeton WordNet (PWN) and its relations, comments on other WordNets, some WordNet extensions and other relevant aspects. One example of an WSD algorithm is UKB[1], which relies on PWN[5] relations which is a way to map context - in a similar way, humans run their own WSD algorithm when faced with a sentence such as "I bought a chair, a couch and a desk.". Because *couch* and *desk* are both types of furniture, it is reasonable to expect that the word *chair* refers to a furniture sense as well, and not a person. A more detailed explanation on how the UKB algorithm works can be found at the introduction of chapter 4.

As put before, an actual NLP pipeline may encompass just some of the previous steps and/or rearrange their order, or it may also combine some steps (like lemmatization and WSD). Although they were briefly described here, each of such steps comprehends a research field of its own and so do the algorithms that combines them. Pondering through some of the many ways this work could contribute to Oil&Gas NLP, we chose to improve a resource instead of an algorithm, i.e., instead of working on a WSD or MWE algorithm or on a full "pipeline", we chose to work on the lexical resource,

[4] https://wordnet.princeton.edu

[5] UKB can also be used with other lexical knowledge base

specifically the WordNet. A lexical resource is a key input of WSD and may support MWE nad NER tasks, among other possibilities, but our focus will be on the WSD impact. The main reason for choosing to improve a lexical resource is to make the best use of the author's strong Oil&Gas background.

1.3 Oil&Gas corpus, an exploratory attempt

This section draws on material previously presented [91]. To conduct an exploratory analysis, we allowed ourselves to briefly deviate from petroleum geology and chose a selection of financial documents. A crucial source of data in for the Oil&Gas business is the collection of electronic documents from the U.S. Securities and Exchange Commission (SEC)[6]. SEC requires companies to file many different forms every year, such as the 10-K form (for US companies) and the 20-F form (for non-US companies). These comprehensive and public documents provide an excellent summary of companies financial and operational performance. They are, however, extensive and intricate - for instance, an integrated company would have information ranging from exploratory efforts up to final products (e.g. diesel) market share, making these documents hard to read even by experienced technical personnel. Companies usually have a dedicated department to produce these documents and to keep up to date with what their competitors are reporting as well.

Given the complexity of the 10-K and 20-F documents, our attempt to build a corpus chose to focus on more straightforward reports. Non-US companies are required to file a Form 6-K whenever information is made public in another country or is distributed to security holders. These are generally short documents with some specific material information - for instance, merge & acquisition operations. In [51], the authors state that foreign companies have generally clearer text than their U.S. counterparts. So initial efforts were employed on the construction of a collection of 6-K forms for a selected group of Oil&Gas companies, namely BP, Equinor, Petrobras, Shell and Total. This group represent some of the most significant listed Oil&Gas companies with activities spread worldwide.

A total of 2,441 forms ranging from 2014 to 2018 was retrieved. Data was retrieved from the SEC website in 'txt' format. But, instead of a plain text format, it is a wrap-up of all the files of which each Form 6-K is made of plus

[6]The EDGAR system is available at https://www.sec.gov/edgar/searchedgar/webusers.htm

a standardized header with all necessary information to identify such filling. In our chosen set any given 'txt' could contain from 1 to 320 files, totaling 10,508 different files. Initial data exploration revealed 'htm' as a core file of many Form 6-K's. Exploratory analysis was narrowed to the forms with at least one 'htm' file: a total of 2,326 forms and 4,695 'htm' files.

Unfortunately, further analysis, which included actual reading many of the retrieved documents showed that there was no standardization among different companies, and even for a given company, report's structure and level of information would vary over time. Also, other than a few simple words like well, platform and alike, there were very few domain-specific words to be retrieved from such corpus. Many had just financial terms, a significant part handled bureaucratic obligations (such as informing changes in the board composition) and very few actually had Oil&Gas signature terms. These reasons reinforced our decision of focusing in geoscience material rather than such financial reports.

Other corpora were procured, but none came through. Scientific papers seemed a great source of technical terms, but with such an extensive domain selecting a representative database alone was a challenge. On top of that, most of available Oil&Gas publications have strong licensing rules, preventing even publication of results based on abstract crawling. Since one of the objectives of this project was to deliver products that could have value for the industry and the research field, any development that could be hindered by licensing liabilities wasn't deemed as a development at all.

Considering these aspects and the previously mentioned issue of representing and balancing all areas and sub-areas of the O&G domain, we chose to use the authority-based approach. This actually reinforced our choice of the subjects lithology and geological time, for it would be challenging to gather documents with good coverage for either but persistent search revealed authoritative sources with comprehensive coverage for both topics.

1.4 Contributions

This project main contributions are:

- Extension of OWN-EN with terms related to lithology and geological time the two main concepts of geosciences

- Extension of OWN-PT with terms related to lithology and geological time the two main concepts of geosciences
- First major use/test of textual approach of OWN-EN and OWN-PT and of the associated mill tool[7]
- Expansion of SUMO to geological time concepts

1.5 Structure

Chapter 2 present Princeton WordNet, its structure and related works on expanding it either vertically or horizontally[8]. It also shows our choice of which WordNets to extend. Chapter 3 presents the lithological domain and our expansion of WordNet onto it as well as onto geological time. Chapter 4 presents our hypothesis and an experiment we ran to test them, as well as some statistics on our contributions to WordNet. Chapter 5 presents the geological time domain and our expansion of SUMO. Finally, Chapter 6 sums up the points covered in previous chapters and points to future works.

[7] section 2.5 briefly describes mill

[8] See Chapter 2 for this proposed classification

Chapter 2

WordNet

This chapter describes Princeton WordNet (PWN) [34], available online at https://wordnet.princeton.edu, where one can use it and access its extensive documentation[1]. The interested reader should consult [42, 59, 60, 41, 33, 89]. It also presents several projects of extending PWN, for which we suggest a classification on types of WordNet extensions: **horizontal** and **vertical** extensions. **Horizontal extension** focus on expanding WordNet to a given language such as Portuguese, Italian, Polish or French, while **vertical extension** focus on extending a given WordNet into a specific domain, like expanding a WordNet for Italian into Juridical domain [85].

Section 2.1 presents PWN structure, relations and highlights the domain issue, while Section 2.2 shows its solution). Sections 2.3 and 2.4 present some horizontal and vertical extensions to WordNet, with the latter concluding that are two ways to do vertical expansion: with a supporting corpus and an authoritative reference material or with just the authoritative reference material - notice that horizontal expansions can be conducted either by corpus or authoritative approach. Also worth noticing is that the translating approach (translating a WordNet in one language into another one) is actually a variant of the authoritative approach where the authority material is a WordNet in another language. Section 2.5 exposes our choice of WordNets to expand, OWN-EN and OWN-PT.

[1] https://wordnet.princeton.edu/documentation

2.1 Princeton WordNet definitions

PWN is a lexical database in and for English. It organizes terms in to four main groups: nouns, verbs, adjectives (which may be subdivided into head synsets and satellite synsets) and adverbs. For each of these groups, words are grouped by synonym sets, the *synsets*. Each synset represents a concept. For instance, the concept *material consisting of the aggregate of minerals like those making up the Earth's crust* is associated with the nouns *rock* and *stone*. The textual definition of the synset is referred to as a gloss, which may also contain examples. The full gloss mentioned above actually is *material consisting of the aggregate of minerals like those making up the Earth's crust; "that mountain is solid rock"; "stone is abundant in New England and there are many quarries"*.

One must not mix the concept with the chain of letters that form words, for words can sometimes represent different concepts. Such words are polysemous words. *Rock*, for instance, can represent the substance as in the previous example or *a lump or mass of hard consolidated mineral matter; "he threw a rock at me"*[2], among several other concepts. Notice the difference between the substance in the first concept and the object in the second concept.

Just like a synset can have one or more words associated, it can also have collocations, i.e., a string of two or more words connected by spaces and/or hyphens. Such is the case of *rock and roll* for the gloss *a genre of popular music originating in the 1950s; a blend of black rhythm-and-blues with white country-and-western*. This same gloss is also associated with *rock 'n' roll*, *rock'n'roll*, *rock-and-roll*, *rock* and *rock music*. As one can see, this is a good example of a synonyms set, a synset: the same gloss is associated with several collocations and a word. The collocations are formed by at least two words connected by spaces or hyphens. We can even detect another example of *rock*'s polysemy.

Synsets are identified by a string of numbers and a letter indicating whether it refers to noun (n), verb (v), adjective (a) or adverb (r). Our previous examples are respectively 14696793-n ("material...") and 09416076-n ("a

[2]Although geologists will criticize the use of the term rock instead of stone, the concept of throwing rocks is so common that some English speaking countries have laws against it, such as New South Wales (Australia) Criminal Act Section 49A: "Throwing rocks and other objects at vehicles and vessels": https://www.legislation.nsw.gov.au/view/html/inforce/current/act-1900-040#sec.49A, last accessed September 30 2020

lump..."). A relevant aspect is that these are PWN identifiers for PWN 3.0. Each version of PWN may have (and usually does have) different identifiers for the same synset. This is due to how they are generated, which considers the number of characters from the beginning of the file to the first character of the synset. Any small change in a synset definition, for instance, alters all the subsequent synset's ids. There are many efforts looking to map and update mappings between ids from different versions of PWN and of other WordNets and other reference materials.

Synsets are organized in lexicographer files: 1 for adverbs, 3 for adjectives, 15 for verbs and 26 for nouns, totaling 45 files. This segmentation reflects syntactic category and logical grouping adopted during PWN's development. For instance, 14696793-n is in noun.substance and 09416076-n in noun.object.

Synsets are linked to each other by means of semantic relations. The most common is hypernym / hyponym[3] relation, indicating a superordinate / subordinate relation, or a "is a type of" relation. For example, 14696793-n is hypernym of *sedimentary rock* and 22 other synsets - see figure 3.1 for a visual. Conversely, *sedimentary rock* is hyponym of 14696793-n. Because a hypernymy relation between A and B implies in a hyponym between B and A (and vice-versa); in PWN only one of such relations is registered in the lexicographer files. On one hand it is a way to prevent file sizes to grow significantly, while on the other hand it helps maintaining an ever-growing resource such as WordNet. Hypernym / hyponym are transitive relations. In our example, *sedimentary rock* is a hyponym of 14696793-n. *Limestone* is a hyponym of *sedimentary rock*. This means *limestone* is a type of *sedimentary rock*, and *sedimentary rock* is a type of 14696793-n. Therefore, due to the transitivity of hypernym/hyponym relations, *limestone* is a type of 14696793-n. Finally, hypernym/hyponym applies to all syntactic categories: nouns, verbs, adjectives and adverbs.

Over the years PWN evolved, encompassing more synsets, word, collocations, relations and types of relations. One can still download many of such previous versions[4]. Version 2.1 was released in March 2005, version 3.0 in December 2006 and the latest version, 3.1, later on. After version 3.1 PWN development came to a halt, and, according to the official website[5], "Due to limited staffing, there are currently no plans for future WordNet releases".

[3]The relations name are underlined the first time their definition is presented, in order to facilitate the reader's navigation.

[4]https://wordnetcode.princeton.edu/oldversions.shtml

[5]https://wordnet.princeton.edu

Aside from the previously mentioned hypernym / hyponym relations, PWN presents several other types of relations. Although optional, relations form the cornerstone of PWN. A relation can link two synsets, in which case is said to be a semantic relation, or it can be a lexical relation, linking to a specific term of a synset. The relations found in PWN were chosen to be intuitively obvious to nonlinguists and to have broad applicability throughout the lexicon [58].

A variant of the hypernym / hyponym relation is the instance hypernym / instance hyponym. It connects a proper noun (the instance hyponym) to a common noun (the instance hypernym) which it is an instance of, like *Obama* is an instance of *President*. Likewise, *President* has instance *Obama*.

Another transitive relation is meronym/holonym, which is actually composed of three sets of relations: member meronym / member holonym; substance meronym / substance holonym; and part meronym / part holonym. If A is a part meronym (or substance meronym or member meronym) of B, then A is a constituent part (or substance or a member) of B. The reflexive relation is B is a part holonym (or substance holonym or member holonym) of A. The part meronym / part holonym applies to nouns that are part of a larger thing, like *arm* is part of *body*. The member meronym / member holonym applies to nouns that are member of a larger thing, like *parent* is a member of *family*. The substance meronym / substance holonym applies to nouns that are made of each other, like a *snowball* is made of *snow*.

Attributes links nouns to their adjectival attributes (and vice-versa), like *beauty* and *beautiful*.

Cause links a verb A that causes a verb B, like *wear down* causes *fatigue*. Cause is not a reciprocal relation.

Entail connects two verbs if the first verb cannot be done unless the second one is or has been done. For example, to *call of* [a scheduled event] entails that something [an event] has been scheduled (from the verb to *schedule*).

Every verb synset contains a list of generic sentence frames to illustrate the types of simple sentences in which the verbs in can be used. In some cases the example sentences are of actual uses of the verbs, but in most cases it is just a collection of generic sentences to point whether the verb accepts something (*a vase breaks*), someone (*a child sleeps*) or just "it"(*it is raining*). There are actually 35 frames in PWN 3.1, and they also point to whether the verbs expect objects, adverbs or other terms[6]. Frames can also

[6]See https://wordnet.princeton.edu/documentation/wninput5wn.

be used to subcategorize verbs, an alternative approach to the traditional transitive/intransitive classification[48].

Adjectives are organized in clusters with a pair (or a triplet[7]) of antonyms called head synsets. Each head synset is linked to other adjectives named satellite adjectives. Similar to is a reciprocal relation linking adjectives. A head synset is always similar to all of its satellite synsets, conversely, a satellite synset is always similar to its head synset.

See also denotes meanings related beyond synonymy and similarity. Originally linked only adjectives, but in PWN 3.1 one of the senses of the verb to *fall* has seven "see also" connections to other verbs. It is a reciprocal relation. One example is *illegal* and *unlawful*.

Same verb group as is a reciprocal and transitive relation between two verbs with similar senses. The transitivity combines groups of overlapping senses into the largest sense groups possible. For instance, to *earn* [money] is in the same verb group as to *make* [money].

Finally, there are the domain of synset / member of this domain relations, subject of the next subsection.

2.2 Semantic Concordance

WordNet originally suffered from what Roger Chaffin called the "tennis problem" [57], [34]. There was no way to map the specialized vocabulary for a domain such as tennis in PWN. Tennis player is in noun.person file, tennis equipment is in noun.artifact, tennis court in noun.location and so on, but there wasn't a connection between such terms.

In 1993 Miller proposed what he called a semantic concordance: a textual corpus and a lexicon so combined that every substantive word in the text is linked to its appropriate sense in the lexicon [58]. His idea was to link PWN to the Brown Corpus[8] [49, 37]. This would give PWN example sentences for many synsets, while syntactically and semantically tagging Brown Corpus words. The main goals were to test and expand PWN's coverage and to develop resources towards what is now known as Word Sense Disambiguation (WSD), but would also point to a possible *topical information* about other

[7] Like solid, liquid, gas

[8] the first major corpus, with over a million words ranging over 15 text categories from political news to book reviews. It is a collection of 1961 texts in [American] English [36]

words that are likely to be used within the discussion of the concept of a synset.

To build such semantic concordance there are two possible approaches: targeted and sequential. Targeted approach starts from a polysemous word in the lexicon, extract all the sentences in the corpus with such word and finally creates a pointer between the lexicon and the corpus. During this process, eventual gaps in the lexicon coverage are fulfilled. This procedure is what lexicographers deem as ideal, for the concentration on a single word at a time provide better definitions. The other strategy starts from the corpus and goes sequentially through it word by word, hence sequential approach. One advantage is it points not only missing senses in the lexicon but also another relevant shortcoming: indistinguishable definitions.

Another example of a semantic concordance is the Princeton Annotated Gloss Corpus[9]. Although incomplete, its goal was to annotate all the glosses (definitions) of PWN synsets. For instance, one of the meanings for *glass* is *a container made of glass for holding liquids while drinking*. The annotation of such gloss would tag container, make, glass, hold, liquid and drink to their respective synsets. Some previous efforts on completing the Princeton Annotated Gloss Corpus within previous versions of PWN were made with both automatic and semi-automatic methods [43], [61]. In [81], we started the manual tagging[10], which was an opportunity to experiment with both targeted and sequential approaches. Given the extent of PWN, the sequential approach was challenging, because for every word the annotators had to understand all the synsets available and then choose the best fit, point a missing sense or overlapping senses. Working with PWN glosses short sentences sequentially revealed to be mentally exhausting. We quickly noticed that recurring words generated recurring doubts and could eventually lead to inconsistent tagging. We then turned to the targeted approach and focused on some domain specific terms. This was much more productive, far less error prone and generated rich discussions on sense picking. It gave us hand on understanding on why it was the chosen strategy for the PWN Gloss Corpus initial phase. The product of this ongoing annotations is used in Chapter 4.

A sort of precursor to the Princeton WordNet Gloss Corpus, eXtended WordNet (XWN) [43] was a project ran at the University of Texas at Dallas up to 2004, and used two different parsers over PWN 2.0 to automatically

[9] https://wordnetcode.princeton.edu/glosstag.shtml

[10] The material is available at https://github.com/own-pt/glosstag

	domain	member of domain			
	noun	noun	adjective	adverb	verb
Topic	434	407	1073	35	1222
Region	166	1199	73	1	2
Usage	29	621	219	72	15

Table 2.1: Topic relations in PWN 3.0

annotate the glosses.

For Semantic Concordance to solve the *topical information* aspect to a given extensive domain such as sports or Oil&Gas, it would need several corpora to be properly tagged, a surmounting task on its own. Even so, the selection and curating of the corpora would have to consider from licensing issues to proper and balanced coverage of subjects, otherwise one could solve the tennis problem but not, say, the baseball problem. So, the eventual solution to the tennis problem came not with semantic concordance itself, but with the introduction of the domain of synset / member of this domain relations, subdivided into three types: topic, region and usage.

Topic refers to the original tennis problem, i.e., a subject such as the game of tennis; region refers to dialectal or spelling variations, such as British English or gangsterspeak; and usage refers to slang - it is actually shown in PWN's graphical interface as in exemplified by / exemplifies, such as *figure of speech* is exemplified by *domino effect*. An example of the domain synset - region relation is *Great Britain* / *call centre* (in US English the spelling would be *call center*). Finally, the most appropriate example for the domain synset - topic relation is *tennis* and 13 words and terms such as *match point*, *service*, *break of serve*, etc.

As [35] points, in Princeton WordNet 2.0 the number of such relations were still very low. According to [39], PWN 3.0 has 440 topics/domains in nouns hierarchy, but actually there are 434 nouns that are domain of synset - topic to 4047 nouns, 1073 adjectives, 35 adverbs and 1222 verbs; 166 nouns which are domain of synset - region to 1199 nouns, 73 adjectives, 1 adverbs and 2 verbs; and 29 nouns that are domain of synset - usage to 621 nouns, 219 adjectives, 72 adverbs and 15 verbs in PWN 3.0. Table 2.1 sums up these numbers.

It is worth highlighting a few points: first, only nouns are domain of synset. Second, there is no restriction to having topic and/or region and/or

usage simultaneously. Finally, a noun can have both "domain of synset" and "member of this domain" relations, like *math* is the domain for *arithmetic* while also a member of the domain *science*.

As [52] states, the use of such relations in PWN seems occasional and without consistent design. Such relations in PWN were annotated by [52] and [17].[11] [52] also endorses the annotation of synsets rather than words or terms and envisages multilingual gains in wordnet-like resources such as EuroWordNet and MultiWordNet [10]. As [39] states, *domain specific resources have proved to be beneficial in tasks such as word sense disambiguation, terminology extraction, machine translation [8], sentiment analysis [24]*, which are more reasons why we chose to extend WordNet to Oil&Gas domain.

2.3 WordNet horizontal extensions

PWN inspired several other similar projects, in English as well as in other languages. To handle such plethora of wordnets, a free, public and non-commercial organization was created: Global WordNet Association (GWA). GWA provides a platform for discussing, sharing and connecting wordnets for all languages in the world[12]. It currently keeps links to 77 different WordNets around the world, covering from Afrikans to Turkish languages, with some multilingual WordNets like the Open Multilingual WordNet[13] [20], which links 34 different WordNets while the Extended Open Multilingual WordNet [19] links WordNets in 150 languages.

GWA focus on WordNets "for all languages in the world", but there are several other variants of WordNets. Section 2.4 covers those which aim to extend PWN to a specific domain like medical jargon.

Most of these initiatives started from some version of PWN and either corrected and extended PWN in English, ventured into some specific domain or translated some (or all) of PWN into another language and eventually expanded afterwards. One relevant exception is PolNet - Polish WordNet[14] [93], [92]: it was manually built from scratch using large scale manual lexicographers work drawing from lexicons, grammars and dictionaries. It is still been updated and expanded, using not only reference material such as

[11] Available for download at http://wndomains.fbk.eu
[12] http://globalwordnet.org/about-gwa/gwa/
[13] http://compling.hss.ntu.edu.sg/omw/
[14] http://ltc.amu.edu.pl/polnet/index.php

dictionaries but also language corpora investigations.

Unfortunately, it is not uncommon to find initiatives that were abandoned along the way. Many papers lead to broken links or projects that haven't been updated in years. There are, however, some ongoing projects that deserve special attention form the reader.

OpenWordNet-EN[15] [63] is an open source project that has converted PWN lexicographer files to textual format, much friendlier to humans - as an example, the hypernym relation is represented by *hyper* instead of PWNs unintuitive @.

OpenWordNet-EN is actually a sister project of OpenWordNet-PT [68], a textual WordNet for Brazilian Portuguese language. It set out to fill the need for an open access WordNet for this specific language. It is free to download, browse online and open to modifications [16].

The English WordNet [53] is an open source project set to be a new WordNet for English. It started as a fork from PWN and in its first release in 2019 corrected over 3,500 minor bugs (such as spelling) and added lemmas endorsed by a manually verified mapping between WordNet and Wikipedia. It has an online interface at https://en-word.net and an open project at GitHub[17]. So far, the English WordNet has lived to its annual release ambitions: 2020 saw a new release [55] with over 15,000 changes over the previous release. [55] shows clear signs of maturing, as the inclusion of guidelines for new synsets and the integration with existing resources such as the Colloquial WordNet [54] (a WordNet for slang) and the Open Multilingual WordNet [20].

In [21] it was proposed the creation of a single shared repository of concepts, the Collaborative InterLingual Index (CILI). CILI is a single list of the union of all synsets of all wordnets Currently, though, it only links different versions of PWN. The project at GitHub[18] hasn't had an update since April 2017.

CILI draws on EuroWordNet's Inter-Lingual-Index (ILI). EuroWordNet [94, 95][19] is a multilingual lexical database linking WordNets for different languages: originally Dutch, Italian, Spanish and English and later extending with German, French, Estonian and Czech. Each WordNet's synset is

[15] https://github.com/own-pt/own-en
[16] https://github.com/own-pt/own-pt
[17] https://github.com/globalwordnet/english-wordnet
[18] https://github.com/globalwordnet/ili
[19] http://projects.illc.uva.nl/EuroWordNet/

linked to the closest (most equivalent) concept in ILI. ILI also connects with PWN version 1.5, an ontology of domains and an ontology of top-concepts. Although the initial seed for ILI was PWN, it has expanded with procedures to add new concepts and relations. This way it is possible to link *cajera* in Spanish and *cassière* in Dutch, even if English doesn't have a word for *female cashier*. ILI registers both theses synsets as synonyms and also links to English *cashier* with a hypernym relation, to indicate that such synsets are a type of a *cashier*, a PWN synset.

2.4 Wordnet vertical extensions

Princeton WordNet (PWN) [34], due to design decisions, does not cover many terms and concepts specific to certain domains as pointed in [23], hence the need to expand PWN for each domain in order to tap into its potential as a NLP resource [5] - this is evidenced by the number of projects for extending WordNets for specific domains are relatively common, a handful of which we cover in this section.

Medical WordNet (MWN) [87] reviews PWN medical terms through a corpus which includes a validated corpus of sentences involving specific medically relevant vocabulary. The corpus is composed by the definitions of medical terms already existing in WordNet, sentences generated via the semantic relations in PWN and sentences derived from online medical information services targeted to consumers. BioWN [78] was another attempt to extend WordNet to the biomedical domain from the Open Biomedical Ontologies (OBO). OBO would provide terms, definitions and relations to be included in WordNet. According to the authors, the attempt failed due to issues on several softwares and resources that eventually prevented the success of the initiative. [23] leans on German's compositional aspect to extend GermaNET with medical terms. The relevance of the candidate terms is then measured in a given domain corpora. Roughly the definitions arise from the compositional rule used to build the term in the first place.

In the legal domain, JurWN [85] builds upon the Italian ItalWordNet (IWN) database, aiming to extend it to the legal domain - IWN [84] is the Italian component of the EuroWordNet [94]. Words were selected from frequent terms used in queries of the major legal information retrieval systems, while definitions were taken from handbooks, dictionaries, legal encyclopedias and other main technical concepts. The LOIS (Lexical Ontologies for

legal Information Sharing) project [77] encompass legal WordNets for six different languages (Italian, Dutch, Portuguese, German, Czech, English) based on the EuroWordNet framework. It used a subset of JurWN as a seed and added new terms on the basis of authoritative resources, national and European Union legislative text and legal text.

GeoName WordNet (GNWN) [18] links the GeoNames[20] geographical database to wordnets in different languages. GeoNames provides both the terms and definitions to be included in GNWN as an instance of a given synset (e.g.: Paris as an instance of city).

Noticeable from all these initiatives is the approach considered to extend a wordnet to a given domain. Some refer to a corpus (custom built or pre-existing material) to gather a list of words to include in a wordnet, and then to an authoritative material such as dictionaries and encyclopedias for the definitions. Others refer to authoritative material that have both terms and definitions, such as ontologies.

Last but not least, the *authority* of the authoritative references mentioned here come not necessarily from an official authority point of view. For instance, one interested in extending a WordNet with culinary terms might choose to use "Grandma recipes" as authoritative source. The *authority* here emanates from the choice itself - in the culinary example, it means that Grandma recipes will be the primary source of terms and definitions, regardless of where it may diverge from, say, Le Cordon Bleu[21].

2.5 WordNet projects to contribute to

We chose to contribute to OpenWordNet-EN[22] (OWN-EN) [63] and Open WordNet-PT[23] (OWN-PT) [68] for a number of reasons.

As [68] states, OWN-PT was created by making a copy of PWN 3.0 and then translating it into Portuguese, while keeping links to the original synsets in PWN 3.0. However, as work progressed, the need to for adjustments in PWN was felt. Since PWN project was halted, the OWN-PT team forked PWN 3.0 once again and established OWN-EN. This means OWN-PT and OWN-EN are sister projects, both with textual approach based in PWN 3.0

[20] https://www.GeoNames.org/
[21] an international chain of culinary schools
[22] https://github.com/own-pt/own-en
[23] https://github.com/own-pt/own-pt

and with structural similarity. Such similarity eased the replication of our OWN-EN extension to OWN-PT, making our extensions both vertical and horizontal. Also, OWN-EN provides a mapping to PWN 3.0[24].

Plus, both OWN-EN and OWN-PT are online open source projects, allowing our contributions to be available to the community. Also, we benefited from proximity with the maintainer, easing updates. Another reason to contribute to both OWNs is to contribute not only to NLP for Oil&Gas in English, the main language for this international industry, but also to contribute for NLP for Oil&Gas in Portuguese - the author works at Petrobras, a major Oil&Gas company and the leading operator in Brazil, with a huge database of documents in Portuguese, internally referred to as a "data mountain".

On top of that, OWN-PT is the textual WordNet for Brazilian Portuguese language chosen by the organizers of the Freeling[25], Open Multilingual Wordnet[26], BabelNet[27] and Google Translate[28] as the representative of the open wordnets in Portuguese used by these projects.

Another relevant reason for OWN-EN and OWN-PT choices is to contribute to these projects as the first major attempt to use their textual format to work on and expand the WN structure. Their projects use mill[29], which performs the same role that grind[30] performed for Princeton WordNet: it processes WordNet lexicographer files in human-readable format, validates syntactic and structural integrity exports to machine-readable formats. With mill one is able to convert between WordNet database format (WNDB)[31], RDF, JSON and textual files. A major difference from grind is that mill allows links between different WordNets, effectively allowing the development of multilingual WordNets in textual format.

[24] Available at https://github.com/own-pt/own-en/blob/master/etc/sensemap.tsv
[25] http://nlp.lsi.upc.edu/freeling/
[26] http://compling.hss.ntu.edu.sg/omw/summx.html
[27] http://babelnet.org/
[28] http://translate.google.com/about/intl/en_ALL/license.html
[29] https://github.com/own-pt/mill
[30] https://wordnet.princeton.edu/documentation/grind1wn
[31] https://wordnet.princeton.edu/documentation/wndb5wn

2.6 Summary

In this Chapter we presented the Princeton WordNet, its structure and concepts. Although its development came to a halt years ago, Princeton WordNet set ground for several others WordNets. The ones that attempted to continue its work on English and/or other languages we classified as horizontal extensions. Other projects, which ventured in expanding a WordNet into a specific domain, where classified as vertical extensions. We presented several of both types of extensions and briefly commented on how they were built. In doing so, it came to our attention that are mainly two ways to do such expansions: with a supporting corpus and an authoritative reference material or with just the authoritative reference material. The corpus is an alternative for selecting words. Either way, the authoritative source always provides the definitions for selected words. We stress that *authority* comes from the choice itself, not from such source being actually canonical in any way or not. Finally, we presented our choice of WordNet projects to contribute to, the OpenWordNet for English (OWN-EN) and the OpenWordNet for Portuguese (OWN-PT).

Chapter 3

Wordnets extension

Part of this chapter's work has been previously published [90]. As stated earlier, among geosciences papers the most common properties raised are usually geographic location [70], geological time and lithological information. This chapter approaches the lithological information aspect and its inclusion into OWN-EN. The inclusion of geological time into OWN-EN and of both topics into OWN-PT is also discussed within this chapter. The geological time aspects are discussed in Chapter 5.

Lithology studies rocks and their formation. It supports geoscientists understanding porosity, permeability, water saturation and other petrophysical properties of rocks which are instrumental aspects of a rock for one working in the Oil&Gas domain. Lithology is described as "the macroscopic nature of the mineral content, grain size, texture and color of rocks"[1]. Therefore, almost any geoscientifical document will at least mention lithological terms, if not resort extensively to such terms.

To the best of our knowledge there is not a comprehensive list of terms and definitions widely accepted by the community. There are some resources which are license-protected like PetroWiki[2] or Schlumberger Oilfield Glossary[3] and some other great offline assets[4]. However challenging, eventually we found a resource that was online, freely accessible and available, suffi-

[1] https://www.glossary.oilfield.slb.com/Terms/l/lithology.aspx, last accessed September 21 2020

[2] https://petrowiki.org/PetroWiki

[3] https://www.glossary.oilfield.slb.com

[4] Many universities, like the University of Tulsa, have thesauri, glossaries or dictionaries dedicated to geological and/or Oil&Gas domain.

ciently comprehensive and structured. Section 3.1 presents such resource and briefly explains its structure.

Section 3.1 present our authoritative source for lithology, section 3.2 presents the actual OWN-EN extension process onto lithology, section 3.3 discusses some decisions, issues and other relevant aspects found during the process, section 3.4 briefly describes OWN-EN extension onto geological time, section 3.5 presents the extension of OWN-PT to both topics and section 3.6 concludes the chapter with some final remarks.

3.1 INSPIRE and GBA authoritative materials

The Infrastructure for Spatial Information in the European Community (INSPIRE) [71] was created to build upon existing resources (infrastructure and data) of the European Union (EU) Member States. The original focus is to support EU policies and activities which may have an impact on the environment. Particularly within the scope of this work, INSPIRE offers an organized codelist for lithology[5]. This resource is actually maintained by the Geological Survey of Austria (Geologische Bundesanstalt) within its "GBA Thesaurus" (GBA). Regarding lithology, GBA presents a richer material than Inspire, all accessible online[6] and available for download[7].

GBA is an ontology based on the Simple Knowledge Organization System (SKOS) vocabulary [46]. Each term has a Universal Resource Identifier (URI) and is related to other terms via SKOS object properties. Within the scope of our work, we have *broader* and its counterpart *narrower*. Therefore, "animal is *broader* than mammal" and "mammal is *narrower* than animal". GBA follows SKOS convention to only assert direct hierarchical links. The name of the term is given by *prefLabel* data property, while the definition is given by *definition* data property. Alternative labels are eventually given by *altLabel*. String values are given in English as well as in German. GBA uses a few other SKOS properties like *related match*, *close match*, *hidden label* and others. Particularly *exact match* is used to map GBA to other resources, INSPIRE included. The downloadable material for GBA is a Resource Description

[5] http://inspire.ec.europa.eu/codelist/LithologyValue
[6] https://thesaurus.geolba.ac.at
[7] https://github.com/schmar00/gba-thesaurus/tree/master/rdf

Framework[8] (RDF) file, which means it is organized in triples consisting of subject, predicate and object.

At its description, GBA states that Lithology comprises loose- and bedrock, classified according to their modal composition and grain size, respectively. Magmatic-, polygenetic-, metamorphic- and fault-rocks are classified based on International Union of Geological Sciences (IUGS) recommendations[9]. Sedimentary rocks classifications refer to international standards. Considering GBA alignment with IUGS recommendations and its mapping to WikiData[10], British Geological Survey (BGS)[11], INSPIRE[12], GeoSciML[13] and DBpedia[14], i.e. several governmental, multinational and community consensual based open-source initiatives, we assumed its Thesaurus for lithology as an authoritative figure. Therefore, it is not scope of this work to question neither the correctness nor the completeness of GBA's material, but rather to map it into the OWN-EN and OWN-PT.

3.2 OWN-EN lithology extension

GBA's labels and definitions of the concepts as well as concepts relations were considered. But GBA's definitions cannot be taken literally since they were not written as dictionary definitions. For instance, they include many in-depth descriptions of the concepts and references to scientific literature. The goal was to provide for the new synsets, as much as possible, Aristotelian definitions[9] following general lexicography methodology.

Besides all information from GBA incorporated into OWN-EN, another product provided is the mapping from GBA concepts URIs to the OWN-EN sense keys. This will also facilitate future revisions of this resource should new releases of GBA be made available. Because GBA is already mapped to multiple other resources (WikiData, BGS, Inspire, GeoSciML and DBpedia), the mapping encompasses these resources as well.

In OWN-EN, the word *rock* has many senses, and the one that resembles the geological meaning is 14696793-n: *material consisting of the aggregate of*

[8] https://www.w3.org/TR/rdf-concepts/
[9] https://www.iugs.org/history, last accessed September 21 2020
[10] https://www.wikidata.org/wiki/Wikidata:Main_Page
[11] http://data.bgs.ac.uk
[12] http://inspire.ec.europa.eu/codelist/LithologyValue
[13] http://resource.geosciml.org/classifier/cgi/lithology
[14] https://wiki.dbpedia.org

minerals like those making up the Earth's crust. The reader should consider this sense wherever *rock* is mentioned henceforth. Figure 3.1 shows how *rock* is represented in OWN-EN, while figure 3.2 shows a few of the uppermost lithologies in GBA. A first look at both shows that WN has at least some hierarchical issues: there are nineteen synsets (in green) that are *hyponym of rock* instead of *hyponym of* one of the three main WN's classes of rock: igneous, metamorphic and sedimentary (all in yellow). Another point of attention is *limestone* (in orange): *hyponym of* both *rock* and *sedimentary rock*. Considering *sedimentary rock* is *hyponym of rock*, the *limestone* to *rock hyponym of* is at least redundant. Finally, an expert would criticize some terms in OWN-EN such as *road metal*, for it is not a proper type of rock, as the definition suggests: *broken rock used for repairing or making roads*.

As shown in yellow in figure 3.2, *sedimentary rock* and *metamorphic rock* are represented in both OWN-EN and GBA. Assessing the definitions, one identifies OWN-EN's *igneous rock* with three counterparts in GBA: *volcanic rock*, *plutonic rock* and *ultramafic rock*. Finally, *limestone* in GBA is *hyponym of carbonate sedimentary rock* which in turn is *hyponym of sedimentary rock*. A relevant point is that GBA does not have a term for "rock" pure and simple. Instead its top concepts are three types of material and from those arise different rocks (deemed "consolidated") and other materials ("unconsolidated", sediments). "Rock" however is used to define other ones (see *sedimentary rock* below). Due to this and to the fact that *rock* is a relevant term in everyday language, we chose to keep this OWN-EN synset, add the three top concepts of GBA and allocate GBA's specific terms downwards from these four synsets.

To expand and adapt OWN-EN onto lithology domain we used GBA's terms and properties starting from the different types of rocks and lithologies. The obvious choice for mapping SKOS relationships to OWN-EN relationships is as first discussed in [11]. In our case, where in GBA A has *broader* B, in WN we defined A as *hyponym of* B; likewise, where in GBA B has *narrower* A, in WN we defined B as *hypernym of* A. For the sake of simplicity, we'll use OWN-EN's relations names henceforth. We also opted for lower case terms when changing or adding synsets in OWN-EN.

GBA does not have explicit relations between rocks and the minerals that compose it, but rock compositions were inferred from GBA's definitions (when possible) and represented in OWN-EN via relations[15] *substance*

[15]See section 2.1 for definitions of such relations

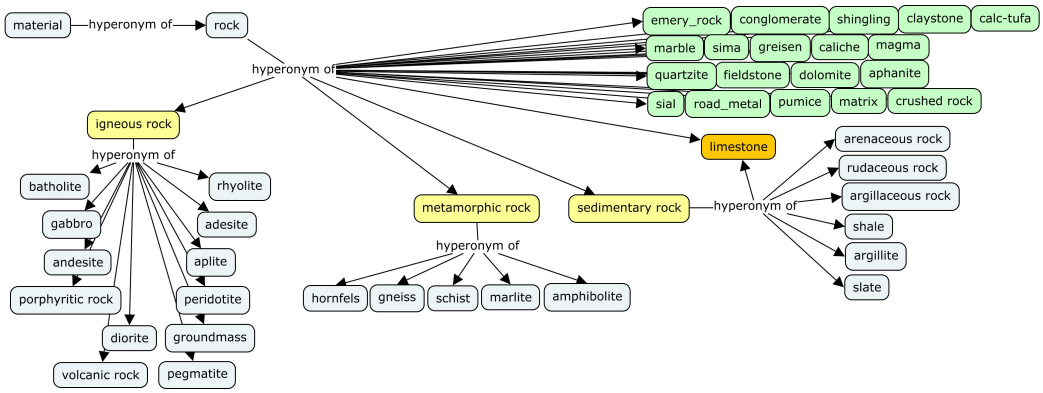


Figure 3.1: Rock in OWN-EN

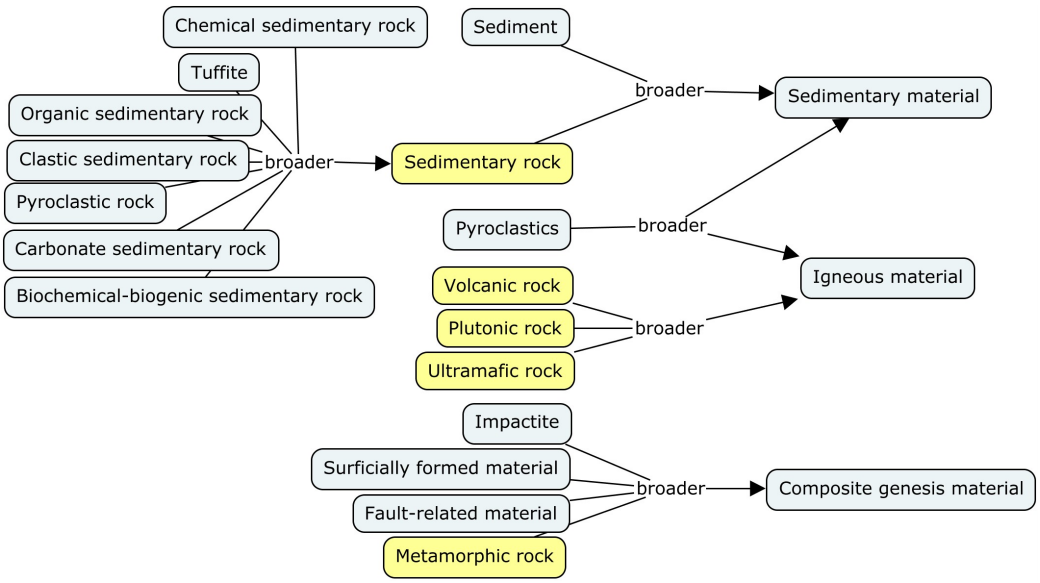


Figure 3.2: Rock in GBA

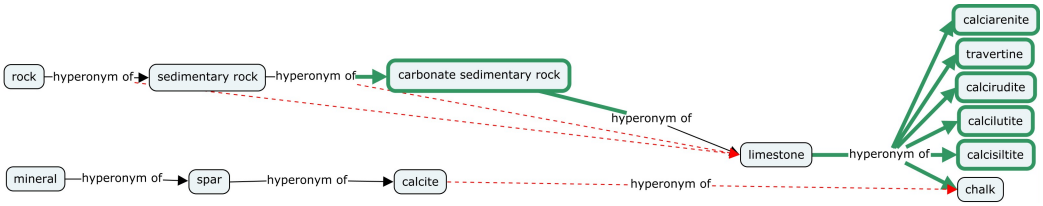


Figure 3.3: Limestone relations in OWN-EN: red ones to be removed, green ones to be included

holonym of and *substance meronym of*. OWN-EN's *domain of synset - TOPIC* and *member of this domain - TOPIC* were also used, as explained later on.

As a proof of concept of this approach, *limestone* was single-picked and initially only definitions and *hypernym of* and *hyponym of* relations were analyzed. Afterwards the *substance holonym of*, *substance meronym of*, *domain of synset - TOPIC* and *member of this domain - TOPIC* relations were included. While the first step enriches OWN-EN with lithological terms, the second step ventures into the mineral domain, expanding OWN-EN even further. Once this work routine was set, the task was expanded to include all *carbonate sedimentary rock* and *clastic sedimentary rock*, the main types or reservoir rocks for Oil & Gas, ergo the most relevant for this industry. We ran a simple test to confirm the potential impact of our contribution thus far, given a positive perspective[16]. We then mapped all of GBA lithology ontology into OWN-EN.

In OWN-EN *limestone* has the aforementioned redundant relations between *rock* and *limestone*. These and other deleted relations are highlighted in red in figure 3.3, where in green is the inclusion of 6 new terms and their 18 new relations with other terms. Note that due to the inclusion of *carbonate sedimentary rock* between *sedimentary rock* and *limestone* the *hypernym of* and *hyponym of* relations between *limestone* and *sedimentary rock* are no longer necessary.

For the six new terms added to OWN-EN we used the GBA definitions with minor adjustments in order to get closer to Aristotelian definitions

[16]Chapter 4 presents an evaluation of both (lithology and geological time) our contributions to OWN-EN, but the interested reader may find a lithology only evaluation in [90]

and general lexicography methodology. For the ones that already existed in OWN-EN, a careful analysis was necessary and carried out top to bottom.

In GBA the concept *sedimentary rock* is defined as *a rock formed from post depositional consolidation of sediments (by processes of compaction, cementation, crystallization, or biogenic binding)* and it is a *hyponym of sedimentary material*[17]. Analyzing both definitions and comparing with OWN-EN's definition for *sedimentary rock*[18] we conclude that, as explained in Section 3.3, OWN-EN's current definition for *sedimentary rock* is technically poor and should be replaced.

Other words OWN-EN already had were *limestone* and *chalk*. *Chalk* was classified as a mineral in OWN-EN, but GBA states that *chalk* is a rock and that rocks are composed of minerals. OWN-EN had synset 14806598-n for chalk: *a soft whitish calcite*, while GBA defines it as *a light-colored (white-gray) marine limestone composed almost entirely of fine crystalline calcite. These porous limestones consist of foraminifera and calcareous algae, and usually contain chert nodules.* On this term we discarded OWN-EN's current definition and replaced it with GBA's.

As for *limestone* OWN-EN has synset 14936226-n: *a sedimentary rock consisting mainly of calcium that was deposited by the remains of marine animals.* The fragment *a sedimentary rock* is represented in the hypernym of relations *limestone* $\rightarrow$ *carbonate sedimentary rock* $\rightarrow$ *sedimentary rock*; the fragment *consisting mainly of calcium* can be addressed by a meronym relation; finally, *that was deposited by the remains of marine animals* is not mentioned by GBA's definition. The first two parts can be removed without losses. As for the last part, [31] states *limestone has two origins: (1) biogenic precipitation from seawater, the primary agents being lime-secreting organisms and foraminifera; and (2) mechanical transport and deposition of preexisting limestones, forming clastic deposits.* Therefore, the whole WN definition for limestone can be disregarded in favor of GBA's[19].

[17] Sedimentary material is defined in GBA as *a naturally-occurring material formed at the Earth's surface, consisting of solid particles aggregated together by one or more depositional processes operating within fluid systems (either aqueous or gaseous) to yield granular particles and/or crystalline particles that are aggregated into layers or bodies. The term includes both unconsolidated sediments and sedimentary rocks.*

[18] In OWN-EN *sedimentary rock* is the synset 14698000-n: *rock formed from consolidated clay sediments*

[19] *a carbonate sedimentary rock composed of* $>$ *95% calcite (and aragonite) and* $<$ *5% dolomite*

Going through the definitions for these ten synsets so far, one can notice three main aspects covered: the process of forming a rock (e.g.: consolidation, compaction, cementation); the constituents of such rock (e.g.: calcite, aragonite); and the size or aspect of the constituents (e.g.: rounded, >2mm). Focusing on the constituents, we confirm that in OWN-EN *rock* is *substance meronym of* mineral, synset 14662574-n *solid homogeneous inorganic substances occurring in nature having a definite chemical composition* in OWN-EN. Reflectively, *mineral* is *substance holonym of rock*.

Combing through the definitions for the nine terms so far under *rock*, we see that the only minerals referenced are *calcite*, *aragonite* and *dolomite*. All three of them already existed in OWN-EN and required only minor changes to the definitions (adding of chemical formulas) and/or to the relations (adding *substance holonym of*) to become adherent to GBA.

Finally, another set of relations was included: the *domain of synset - TOPIC* and *member of this domain - TOPIC*. Given the topic of choice, all of the terms added from GBA's lithological terms were associated with *lithology* domain and their constituents with the *mineral* domain.

The *limestone* example is where we designed and field-tested our approach to map GBA into OWN-EN. Six new terms were included and four previously existent synsets corrected, along with their *hypernym of* and *hyponym of* relations. As we analyzed *substance holonym of* and *substance meronym of* relations, some of GBA's mineral terms were included in OWN-EN. It is not the scope of this work to cover all of GBA's minerals, but the ones mentioned in the rock's definitions were included.

Following this same approach, we were able to include all of *carbonate sedimentary rock* and *clastic sedimentary rock*, encompassing 27 new synsets with new 79 relations and 9 definitions changes, 15 removed relations and 71 new relations in pre-existing synsets. These types of sedimentary rocks represent the two main types of oil & gas reservoirs throughout the world[47]. Finally, for the sake of completion, we included all the remaining GBA lithological terms into OWN-EN.

3.3 Discussions

The extension of OWN-EN raised some relevant points. In this section we cover such points and explain the reasoning behind the decisions made within the possibilities considered.

A recurring matter is regarding the multiword expression (MWE) issue. Should we keep and create a synset for a MWE? Or is it enough to have all words individually in the resource? For instance, in OWN-EN we have a synset for *sedimentary rock*[20], but is it a *rock*[21] that is *sedimentary*[22]? In other words, is *sedimentary rock* a necessary synset or could one deprehend its concepts from the synsets *sedimentary* and *rock*? Likewise, GBA subdivides *sandstone*, *sand*, *siltstone*, *silt* and *gravel* into *fine*, *medium* and *coarse*, meaning *fine* presents more and smaller grains than *medium* which in turn has more and smaller grains than *coarse*. But GBA sets a specific grain diameter range for *fine sandstone* which is different from the range of *fine siltstone* (respectively 0.063mm to 0.200mm and 0.0020mm to 0.0063mm). Due to this aspect, one possibility would be to adjust existing (or create new) synsets to ensure that *fine*, *medium* and *coarse* retain their relative properties, but the cutoff values (e.g.:0.063mm to 0.200mm) would be lost. In such cases we chose to respect our authoritative source, not just for being technically [more] accurate, but also to leave a potential bridgehead to other languages - a given sense may or may not have a proper word form in other languages. For instance, while in English we have *marl* and *marlstone* for a given sediment (unconsolidated) and its counterpart rock (consolidated), in Portuguese both can be *marga*[23]. This argument also favors registering *fine siltstone* as a synset, i.e., although in English one could see it as a *siltstone* which is *fine*, in other languages this sense may have a word. The existence of the *fine siltstone* synset should ease the process of mapping WN for different languages, hence improving a translation task, for instance. Section 3.5 continues the cross-language aspects of this discussion from the Portuguese and translation points of view.

Another issue we faced was when layman's knowledge clashes with technical definitions. For instance, for *sedimentary rock* the synset 14698000-n definition reads *rock formed from consolidated clay sediments*. From a technical perspective, clay is an unconsolidated sediment with very small grain, whilst *sedimentary rock* can be formed from several grain sizes, so we re-

[20] synset 14698000-n *rock formed from consolidated clay sediments*

[21] synset 14696793-n *material consisting of the aggregate of minerals like those making up the Earth's crust*

[22] synset 02952109-a: *resembling or containing or formed by the accumulation of sediment*

[23] *Marlstone* can also be translated to *pedra marga*.

placed OWN-EN's definition with GBA's[24]. Another example is sandstone, synset 14995541-n: *a sedimentary rock consisting of sand consolidated with some cement (clay or quartz etc.)*. Even though OWN-EN definition was not so far off, it presented *sandstone* as an *hyponym of arenaceous rock*, synset 14697485-n: *a sedimentary rock composed of sand*, a term not present in GBA. On the technical side *sand* is a clastic sediment within a certain grain size range, but OWN-EN defines sand as being silica-based, i.e., the sand commonly found in beaches, a common misunderstanding even among technicians. In order to accommodate such divergent points, we merged *arenaceous rock* and *sandstone* synsets, kept the seven synsets *sandstone* was already *hypernym of* and then complemented with GBA's material.

3.4 OWN-EN extension to GeoTimes

Geological time is described in detail in Chapter 5, where we also describe the authoritative source: ISC ontology [27], an OWL ontology for the geologic timescale. In this section we attain to comment on OWN-EN extension itself.

As [63] discuss, Princeton WordNet only presents "28 synsets dedicated to the most well-known geological periods", many of them with outdated senses and terms according to ICS more recent versions [25][25]. In order to allow processing of old texts which used the outdated terms, we kept such synsets and adjusted their definitions expliciting it incorrectness (e.g. *Tertiary* had its definition set to *an obsolete term for the geochronologic period ranging from 66 millions of years ago to 2.6 millions of years ago; obsolete term that refers to Paleogene and Neogene geochronologic period*).

The mapping of the totality of such terms is rather straightforward, specially after the know-how acquired during the lithology extension. We chose to keep definitions simple, using the frame "a geochronologic age / period / etc ranging from X millions of years ago to Y millions of years ago". For most of the 28 synsets previously present in WN, we squashed this frame with the previous definition, which was rather the same frame with inaccurate boundaries and/or some characteristic of such time, which was added with a "distinguished by..." preamble. For example, for the Jurassic, we have "a geochronologic period ranging from 201.3+/-0.2 millions of years ago to

[24] *a rock formed from post depositional consolidation of sediments (by processes of compaction, cementation, crystallization or biogenic binding)*

[25] https://stratigraphy.org/chart

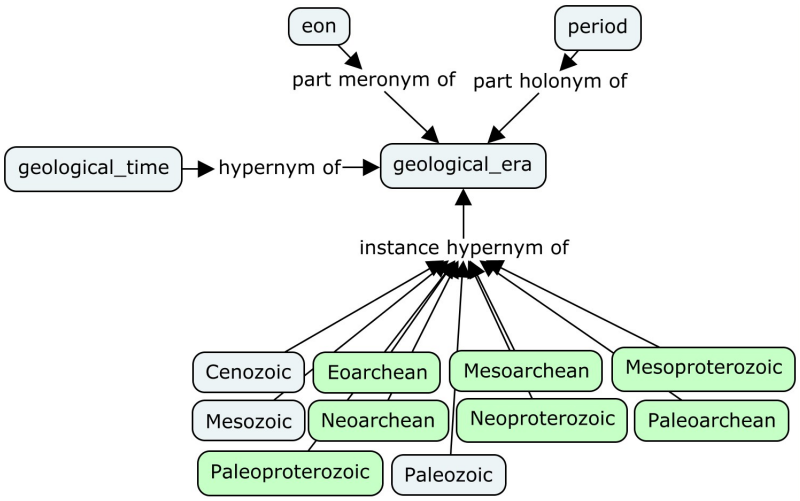


Figure 3.4: Era in OWN-EN, in green our inclusions

circa 145.0 millions of years ago; distinguished by the presence of dinosaurs and conifers; distinguished by the appearance of first birds".

As for the relations, every age, period, etc was deemed as an instance hypernym of its type. We also connected each unit with its building blocks immediately under and above with the meronym part ("mp") and holonym part ("hp") relations. Figure 3.4 depicts changes in *Era* relations, while figures 3.5 and 3.6 shows, respectively, how *Jurassic* was represented in the original OWN-EN and after our contributions. The full synset for *Jurassic* is:

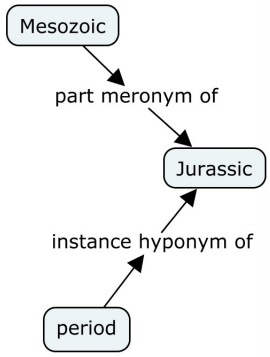


Figure 3.5: Jurassic as originally in OWN-EN

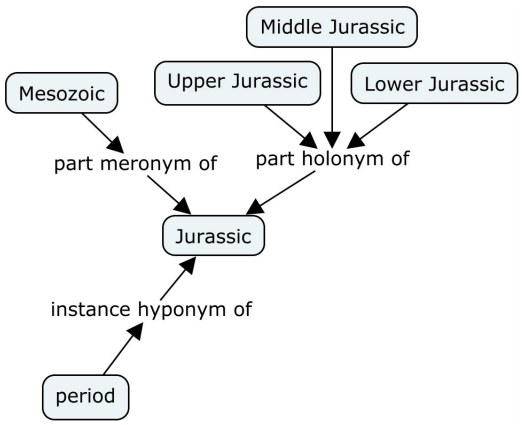


Figure 3.6: Jurassic in OWN-EN after our contributions

```
w: Jurassic
w: Jurassic_period
d: a geochronologic period  ranging from 201.3+/-0.2
   millions of years ago to circa 145.0 millions of years
   ago; distinguished by the presence of dinosaurs and
   conifers; distinguished by the appearance of first
   birds
hp: Mesozoic
ihyper: geochronologic_period
mp: Lower_Jurassic
mp: Middle_Jurassic
mp: Upper_Jurassic
```

Notice, however, the limitations of the lexical resource: although one can infer that Jurassic is composed of three parts, there is no information about their order or [lack of] overlapping. Naturally one could read all three definitions and come to the conclusion that these are sequenced parts whose beginning and ending match perfectly each other and the first of such beginnings and last of the ending match those of the Jurassic itself. Even so, one could still argue whether it is possible or not to have Lower Jurassic starting at 201+0.2 millions of years ago and the Jurassic itself beginning at 201-0.2 millions of years ago. Wordnet simply lacks the resources to enforce this time logic concepts in a reasonable way. The semantics are left to the imagination of the user (and each user is likely to have a slightly different intuition),

rather than accessible through logical inference. Chapter 5 addresses this kind of concern.

WordNet has some other limitations. For instance, throughout its versions, ICS altered some names and some boundaries of several geological time entities. Although one can add all possible terms to a given sense, it is not trivial to point which ones are still valid or not. The same goes to relations: there is no simple way to point expired ones and current ones. As an example of the last one, consider Texas: it is a state, hence instance hypernym of state. But for almost a decade it was an independent country, therefore it should also have an instance hypernym relation to country, but with some sort of tag stating its only valid between 1836 and 1846. WordNet cannot represent such an intricate concept.

3.5 OWN-PT extensions

Once our work with extending OWN-EN vertically onto lithology and geological times for English was done, we used it to support the horizontal extension in Portuguese in OWN-PT. Considering the structure of relations was already done in OWN-EN, this should be only a matter of procuring proper translations. Regarding geological time it was indeed as straightforward, given ICS has official translations to several languages, Portuguese included. The fact that we chose a standard frame for definitions facilitated the translation, leaving only one minor issue with the eventual extra information: some technical names for lifeforms in the "distinguished by" portion of the definition, like in Ordovician "distinguished by the appearance of conodonts and ostracods". Wikipedia and its articles for different languages provided an easy solution.

But the actual unforeseen challenge came with lithological terms. Our extension in English was based in GBA, which only comes in English and German. GBA is, however linked to INSPIRE, which is an European Union (EU) initiative and should have all terms in EU languages, Portuguese included. Unfortunately, it only had translations for very few terms and definitions, 31 to be exact. Even so, we had to put in some work to find common ground definitions between INSPIRE in Portuguese and OWN-EN. For all the other synsets we had to manually translate all terms and definitions. The author's Oil&Gas background was instrumental in this process.

It was during the Portuguese extension we came across cases like the

marga example mentioned in section 3.3. We also faced several concepts which hadn't a proper translation, such as *quartz alkali feldspar syenite*[26] for EN / *sienito com quartzo e feldspato alcalino*[27] for PT. The wording in Portuguese doesn't form what one may acknowledge as a proper MWE[28] but we chose to be consistent with the previous decision taken during the EN expansion and keep the term as an MWE. Not only it respects our authoritative source[29] for the term translation, but it is also the bridgehead on the Portuguese side.

Finally, with the so-called bridgeheads on both sides, we were able to add a relation "same as" to cross reference between OWNs. This kind of relation allows, for instance, one checking the Portuguese word *brecha piroclástica* to immediately link it to its English counterpart *pyroclastic breccia*, as in the example below (taken from the extend version of OWN-PT):

```
w: brecha_piroclástica
d: uma rocha piroclástica com mais de 75% de
   blocos angulares (diâmetro >64mm)
hypo: rocha_piroclástica
mt: noun.cognition:litologia
sa: @en:noun.substance:pyroclastic_breccia
```

3.6 Summary

In this chapter we procured an authoritative source for lithology related terms, found one (GBA) which was structured on SKOS. Unlike geological time, there is no single authoritative source for lithology that could be found

[26] *a plutonic rock and variety of syenite in which plagioclase comprises <10% of the total feldspar*

[27] Although this exact wording is used in literature, such as [88], this is not the prescriptive translation according to IUGS standards and guidelines. The prescriptive term is "quartzo feldspato alcalino sienito" (as in [29]); given the reach of INSPIRE initiative, we chose to keep the descriptive terms, i.e., INSPIRE choice of Portuguese translations, and suggest that prescriptive names be added to OWN-PT in future works.

[28] The equivalent in a free translation to English would be "syenite with quartz and alkaline feldspar".

[29] http://inspire.ec.europa.eu/codelist/LithologyValue/quartzAlkaliFeldsparSyenite

for use in this project[30]. Although one may question the completeness of GBA, it nevertheless provided the basis for generating what might be critical mass of lithological terms in OWN-EN - as suggested by the relevant impact on better disambiguation within preexisting synsets in the experiment ran in Chapter 4. As the Portuguese example of *sienito com quartzo e feldspato alcalino* in Section 3.5 showed, some terms might not be adherent to prescriptive rules and conventions, but function as descriptive terms, i.e., are used in the literature. It is a future work asses the need and to include prescriptive terms in OWN-EN and/or OWN-PT.

We then presented how the WordNet extension process was organically designed, field-tested, and conducted up to its completion. We discussed some decisions, issues and other relevant aspects which came to our attention during such process for the lithology extension and the geological time extension in English and in Portuguese. We were also able to include some of GBA material on minerals, laying ground for future works to include all of it in OWN-EN.

This chapter also showed that when dealing with technical terms which might be closer to everyday English, one risks facing words and/or senses already in OWN-EN. This poses some challenges like in the *sand* and *limestone* cases, where we had to decide on whether to merge the previous OWN-EN definition with the technical one, keep both or just one definition, and if so which one to keep. Nouns that were not previously in OWN-EN and/or with only one sense even in outside of OWN-EN, like *porphyroid*[31] did not pose such challenge, but demanded a careful analysis on the definition (whether and how to adapt it) and the relations - how to map the ones in the authoritative source with the ones available in WordNet, how to handle ones that don't show a clear correspondence, etc.

[30]There are several respected sources for lithology, but most of them adopt licenses that prevent using in a project such as this one; and/or are only available in media or data format which would encumber and possibly render this project not viable

[31]*metamorphosed volcanic rocks of usually acid to intermediate chemistry (ignimbrite, tuff, tuffite) with porphyroblasts in a fine-grained groundmass*

Chapter 4

Evaluation

In order to assess the impact of the aforementioned extensions of WordNet, we will focus on Word Sense Disambiguation (WSD), a concept briefly described in Section 1.2. One recurring method of conducting evaluations in WSD is comparing results over a golden set, i.e., a corpus annotated with at least the correct PoS and senses. Although this a relevant way of assessing an algorithm performance, it is not a proper way to assess changes in a resource such as WordNet - since we have two versions of WordNet, we would need/create two golden sets, but with no way to assess our WordNet contributions. Instead, our hypothesis is that a tool which relies on WordNet has an overall performance better with the extended OWN-EN than with the original OWN-EN, more precisely:

- new words: words with no previous sense should be properly tagged
- preexisting words with new senses: words previously without proper sense should be properly tagged
- preexisting words for preexisting senses: words which had a proper sense but were misstagged might (should?) be properly tagged

Notice that first part of our hypothesis is rather patent: words which didn't exist in OWN-EN and were added should be properly detected and tagged, regardless of whether they were associated with new or previously existing senses. The second part lies within reasonable expectation: if the WSD algorithm only had inappropriate senses to choose form and now has a correct one, it should pick properly.

The final part of the hypothesis comprises our scalability expectations. By adding not only new words and senses but also new relations, we expect the WSD algorithm to exploit such relations and choose better on words and senses which had little to no changes in their synsets. This better disambiguation should happen due to not only the relations of the word being disambiguated but also to the relations of other words within the same sentence.

Our experiment uses the corpus studied by [80]. It consists of over five thousand sentences, with an average 28 words per sentence. It was built from 1298 publicly available English language geological reports, published by the United States Geological Survey, Geological Survey of Canada and British Geological Survey.

4.1 UKB Algorithm

One such algorithm that notably exploits WordNet relations is UKB[1]. UKB builds a graph G with WordNet concepts as nodes and WordNet relations as undirected edges. From an input text, UKB identifies content words (nouns, verbs, adverbs and adjectives) present in WN and, for each of such words W_i, creates a list with all m concepts possible $Concepts_i = \{v_1, ..., v_m\}$. It then adds the W_i words to G, linking them to their respective $Concepts_i$. Finally, for each word to be disambiguated (W_i), it uses the PageRank algorithm [22] to rank the $Concepts_i$ nodes for each W_i, hence completing the word sense disambiguation process. To rank the nodes, PageRank considers that every edge from v_1 to v_2 is a vote from node i to node j, which increases the rank of j proportionally to i's rank. The bootstrap is done by assigning equal ranks uniformly over all W_i, which then influences the rank of all $Concepts_i$, spreading their mass over G over iterations. The algorithm stops when it reaches either a given limit number of iterations or convergence within a given threshold.

There are actually two ways to implement PageRank within UKB. One option is the Personalized PageRank, where G is built with the full WordNet. The other option is Traditional PageRank over Subgraph, where instead of G it relies on G_D, a subgraph of G with only the nodes of all $Concepts_i$ and the edges which are the shortest paths between each node of $Concepts_i$ and $Concepts_j$ for every pair i, j. In other words, G_D represents the UKB hypothesis of capturing the most relevant concepts and relations from WordNet

for a given input context. For example, in "Chloe likes pop music and rock", *rock* is most likely to be tagged with the *rock music* sense due to its proximity with the *pop music* sense, since both are connected to *popular music genre*, while their other senses are further apart. In other words, it is more likely that the sentence is about two genres of music that Chloe likes than about how she likes music and stones.

As expected UKB doesn't accept raw text as input, instead it expects a certain degree of preprocessing. For this task we chose Freeling [67], described in the next section. Freeling not only preprocess the texts but it also implements UKB.

4.2 Freeling library

Freeling [67] is an open source language analysis tool suite that not only does this preprocessing but also implements UKB within its library. Within such library one can customize which modules should be called, and for our experiment we called:

- tokenizer: transforms plain text string to a list of tokens. For instance, it recognizes abbreviations (e.g., i.e., etc.), e-mail addresses, #hashtags and alike (other regular expressions) as a single token;
- sentence splitter: in a list of words, detects sentences boundaries and returns a list of sentences. It is able to work with sentence markers unambiguous like "?" (always marks a sentence ending) and ambiguous like "." (may or may not mark a sentence ending);
- morphological analyzer: like analyzer, it is not actually a module but rather a metamodule that calls several submodules:
  - Punctuation Detection: assign a PoS tags each punctuation symbol;
  - Number Detection: assigns a normalized value as lemma to numerical expression like 2,179 and fifty-four;
  - User Map Module: assigns lemma and PoS to a set of specific strings or regular expressions, like <HTML>;
  - Dates Detection: for each date expression, sets it as a multiword expression and a standardized lemma;

  - Dictionary Search: for words in the dictionary, recovers their lemmas and PoS tags. If the word form is not within the dictionary, it applies affixation or compounding rules to retrieve the same information;
  - Multiword Recognition: aggregates multiword tokens in a single token, provided they match a given list of MWE;
  - Named Entity Recognition: recognizes NE, primarily by identifying capitalized words;
  - Quantity Recognition: identify ratios, percentages and physical or currency magnitudes such as 20%, 2 pounds and twenty dollars. It then converts it to standardized lemma and PoS tag;
  - Probability Assignment and Guesser: if no PoS tag is assigned so far, it guesses one based on word ending. Also, to each analysis of each word, it assigns an a priori probability which will be later used by the PoS tagger

- PoS tagger: perform PoS tagging;
- sense annotation: for each lemma, returns all available senses from the sense dictionary;
- WSD: runs UKB algorithm to rank the possible senses;
- NEC: classifies Named Entities into person, geographical location, organization and others;
- parsing: adds to each sentence a parse tree and then converts it to a dependency tree and, for each edge in it, assigns a syntactic function;
- SRL: performs the semantic role labelling;
- coreferences: enriches a document metadata with coreference information. Since coreference is document-level aspect, it does not process single sentences or paragraphs.

4.3 Limitations

It is important to notice that, despite delivering comparable results, the choices above come with their own limitations and implications. UKB, for

instance, is one among many WSD algorithms and one could argue in favour of other candidates such as the Lesk Algorithm [50][1]. Freeling handles all the processing from the corpus to the WSD result, but has its own limitations. For instance, it only recognizes MWE with adjacent terms, meaning it cannot recognize *Albert Einstein* in line such as Albert the genius Einstein [2]. It is, as other NLP suites, subject to cascading effetcs: an error in a preliminary step, such as tokenization or sentence splitting, could cause havoc in the WSD results. The mere order of such steps can produce different outputs: a PoS tagging prior to WSD may negatively affect the later. As for the chosen corpus, as stated in Chapter 1, any corpus for the Oil&Gas domain will have sub-area representing and balancing issues, and as put in Section 1.3, most Oil&Gas licensing rules are so strict that even publication of results based on abstract crawling is forbidden.

All of these points considered, the propose of this whole experiment is to serve as a proof of concept for our hypothesis, not to exhaust all possible effects of its design. Hence the chosen corpus was available and is geology-orientated, making it a reasonable candidate for testing our hypothesis, while UKB and Freeling are as good a choice as any for our purposes here.

4.4 WordNets to compare

Our starting point is the PWN version 3.0, or, more accurately, OWN-EN textual version of PWN 3.0. We ran our experiment with this version of OWN-EN and then again with the current version of OWN-EN, which is essentially the original version after all of our contributions. We will refer to the initial version of OWN-EN as *original OWN-EN* and to the later as *current OWN-EN*.

Original OWN-EN had 155287 words (lexical forms), 117659 senses (synsets), 206978 pairs word-senses and 205641 relations. Our contributions removed 10 words, 1 sense, 10 pairs word-senses and 53 relations, while adding 557 words, 364 senses, 567 pairs word-senses and 1827 relations. Current OWN-EN has 155834 words, 118022 senses, 207535 pairs word-senses

[1] Lesk compares the possible definitions for given word with the ones from neighbouring words, the definition with most words in common with its neighbors is picked. One implementation with WordNet [13] also uses the definitions of senses related by a given set of relations such as hypernym/hyponym.

[2] Non-contiguous MWE is still a relevant issue for NLP[14]

	word	sense	word-sense	relation
Original	155287	117659	206978	205641
removed	-10	-1	-10	-53
added	+557	+364	+567	+1827
Current	155834	118022	207535	207415

Table 4.1: Effects of our contributions on OWN-EN

and 207415 relations. Table 4.1 summarizes this numbers. For instance, consider the example of the word *tuff*: in original OWN-EN its synset had 2 words (*tuff* and *tufa*), 1 sense, 2 pairs word-sense and 1 relation (hyponym of *volcanic rock*); afterwards, it came to 3 words (added *ash tuff*), 1 sense (with a revised definition), 3 pairs word-senses and 2 relations (1 deleted and 2 added).

```
Original OWN-EN
w: tufa
w: tuff
d: hard volcanic rock composed of compacted volcanic ash
hypo: volcanic rock

Current OWN-EN
w: ash tuff
w: tufa
w: tuff
d: pyroclastic rock with more than 75% ash (diameter <2mm)
hypo: pyroclastic rock
mt: noun.cognition:lithology
```

4.5 Freeling input files

To actually use Freeling, we wrote a script[3] which inputs OWN-EN files and outputs Freeling files. The impact of WordNet changes is represented in four Freeling inputs:

- dicc.src: a dictionary to associate different variations of a word to its lemma and PoS (e.g. *rocks* might be related to the noun *rock* or to the

[3]available at https://github.com/own-pt/own-en

verb *rock*)

- locucions.dat: identifies MWE, their lemmas and PoS (e.g. *sedimentary rock*)
- senses30.src: associates a synset_id to its respective set of synonyms (e.g. 14696793-n to *rock* and *stone*)
- wn.dat and xwn.dat: indicates a relation between two synsets (e.g. 14696793-n 14698000-n to point a relation between *rock* and *sedimentary rock*)

A point of attention is the last file: while the original Freeling wn.dat file holds the relations (hypernym, holonym, domain, etc), xwn.dat encompasses wn.dat and the relations of eXtended WordNet[4] - a project ran at the University of Texas at Dallas up to 2004, which used two different parsers over PWN 2.0 to automatically annotate the glosses. It is a sort of precursor of the Princeton WordNet Gloss Corpus - as mentioned before it is an initiative which has not been concluded, but shows enormous value, as similar attempts attest [43, 61, 81].

In our context, we generated a wn.dat and a xwn.dat file for each OWN-EN version (original and current). While each wn.dat file represents only the actual relations registered in each OWN-EN version, our xwn.dat file uses our work from [81].

4.6 Experiment

We ran the experiment 4 times varying the OWN-EN version (original or current) and the relations file (wn or xwn). Since WordNet only has senses for adverbs, adjectives, verbs and nouns, we focus our analysis only on such *content words*: with original OWN-EN there were 96590 content words, and with current OWN-EN 96247 content words. The difference of 343 terms is due to the new MWEs added to current OWN-EN such as *Lower Cretaceous*: where previously there were 2 terms now there is only 1 term, a MWE. Table 4.2 shows the number of synsets that changed when altering just the relations file. Several sentences presented no alteration in any scenario, like the sentence below:

[4] https://github.com/TALP-UPC/FreeLing/issues/105#issuecomment-684507586

	wn file	xwn file
# of changes	1103	1184

Table 4.2: Number of content words which changed synsets

	Orig. OWN-EN	Curr. OWN-EN	difference
wn	219230	221005	+1775
xwn	673537	675310	+1773
difference	+454307	+454305	

Table 4.3: Relations per relations file (wn/xwn) and OWN-EN version

(1) Most workers agree that the upper contact of the Travis Peakwith overlying shallow-marine carbonates of the *Lower Cretaceous* Sligo Formation is conformable.

As shown in Table 4.3, the use of xwn relations file instead of wn provides over 45 thousand new relations for both original and current OWN-EN. On the other hand, our contributions provides under two thousand relations. Since our WSD algorithm[5] draws upon the relations, we focus the next section on the results with the xwn file. After all, our contributions affected less than 1% of total relations, but the choice of the relations file causes a threefold increase in the number of relations, which directly affects the performance of UKB.

4.7 Commented results

This Section highlights terms which were directly in our scope, i.e., lithology and geological time terms and also terms relevant to the Oil&Gas domain. It focus on comparing results solely with xwn file.

The noun "age" had 158 occurrences in our corpus, 138 of which were tagged to synset 15153787-n: *a late time of life*. With Current OWN-EN 150 changed to age as in *geological age*, 7 changed to synsets associated to MWEs, namely *Albian age*, *Maastrichtian age* and *Santonian age* and only one did not change synset.

Other 56 of the new synsets for geological times were properly tagged 475 times, none of which had synsets previously associated. For instance, *Lower*

[5]see section 4.1

Cretaceous and *Campanian* were two of such new synsets, respectively with 90 and 24 occurrences.

Within the new synsets for lithological terms, there were 87 occurrences for 16 such synsets, none of which had synsets previously associated. *Anhydrite* and *black shale* being the most frequent, with respectively 22 and 18 occurrences.

Of the 22 occurrences of *anhydrite*, 4 were tagged as verb, 12 as adjective and only 6 as nouns with original OWN-EN, although none had a sense tagged. With current OWN-EN all were tagged as nouns and to the new sense *an evaporite composed of >50% anhydrite (CaSO4); often associated with gypsum and halite, forming thick interbedded layers with limestone and shale*. Bellow one such example where *anhydrite* occurs twice, and was once deemed as an adjective and another as verb:

(2) Maximum hydrocarbon column in the Anschutz Ranch East trap is 2,1 ft. Major seals include **anhydrite**[6] in the Twin Creek limestone and salt in the Preuss Sandstone, where present, overlying Nugget oil reservoirs in the eastern fold trend, and **anhydrite**[7] at the top of the Madison Group, along with thick shale in the Triassic section capping wet gas Paleozoic reservoirs in the two western lines of folds.

Dolostone is one example of new words within preexisting senses: previously without any synset tagged by the WSD algorithm, it was later properly tagged 13 times, such as in the example below.

(3) Known reservoir rocks include lateral equivalents of the lower Eocene Bou Dabbous Formation, such as the El Garia fractured limestone; Eocene Jirani **dolostone**, Jdeir limestone, and Reineche limestone; Oligocene to Miocene Ketatna limestone; and the middle Miocene Aïn Grab limestone and Birsa sandstone.

Sediment is a word that previously had a only one noun sense: 09428967-n *matter that has been deposited by some natural process*, registered in the noun.object lexicographer file. With original OWN-EN this was the tagged synset for all 109 occurrences. With current OWN-EN 11 became part of

[6]With Original OWN-EN this occurrence was deemed an adjective, while with Current OWN-EN it was properly tagged with the correct noun synset

[7]With Original OWN-EN this occurrence was deemed a verb, while with Current OWN-EN it was properly tagged with the correct noun synset

MWEs (9 *clastic sediment* and 2 *carbonate sediment*), 10 did not changed synset but the remaining 88 were tagged with the new synset 90000188-n *solid fragmental material that originates from the weathering of rocks and is transported or deposited by air/water/ice, or accumulates by other natural agents such as chemical precipitation from solution or secretion by organisms; forms in layers on the earth's surface, in a loose or unconsolidated form.*

With original OWN-EN, *Cretaceous* was deemed an adjective 540 times, 9 of which tagged as 03057732-a *abounding in chalk* and 531 as 03057591-a *of or relating to or denoting the third period of the Mesozoic era.* With current OWN-EN, Cretaceous became a part of new MWEs *Lower Cretaceous* in 90 occurrences and *Upper Cretaceous* in 86 occurrences, kept the 03057591-a tagging 363 times and reduced the 03057732-a to only 1 occurrence.

Within the group of previously existing words and senses, several did not change synset. Within the Oil&Gas domain, *basin*, *source* and *well* stand out. The predominant model for Oil&Gas all over the world considers that all hydrocarbon accumulations occurs within a sedimentary basin. In our corpus, *basin* has 1282 occurrences, most (1254) were incorrectly tagged as 02801525-n *a bowl-shaped vessel; usually used for holding food or liquids.* *Source* is a frequent term (with 488 occurrences in our corpus) to explain the origin of hydrocarbons, most (364) of which were improperly tagged as 06675122-n *a document (or organization) from which information is obtained.* The noun *well*, as expected a frequent term (415 occurrences), was tagged all times as 06675979-n *an abundant source.*

Another frequent word in Oil&Gas domain is *formation*: it occurs 1012 times in our corpus. With original OWN-EN it was properly tagged as 09287968-n *(geology) the geological features of the earth* 56 times while misstagging the remaining ones. With current OWN-EN mistagged another 4 occurrences.

Rock was properly tagged 137 times with synset 14696793-n *material consisting of the aggregate of minerals like those making up the Earth's crust* with original OWN-EN, but with current OWN-EN this number went down to 105 times - most of the previous tagging was relocated to synset 10535366-n: *(figurative) someone who is strong and stable and dependable.* Below is one example of such case:

(4) A subsequent minor marine transgression is recorded by the Gilmer Limestone ("Cotton Valley limestone") in east Texas, although equivalent facies in northern Louisiana and Mississippi are terrigenous clastic

rocks known as the Haynesville Formation.

Interestingly, these examples of *formation*, *source*, *rock* and *basin* fall into the cases of "one sense per discourse"[38]: well-written discourses tend to avoid multiple senses of a polysemous word. This actually applies to polysemous MWE [96] and entities [15] as well. We leave as a future wok one interesting exercise: to rerun this experiment while forcing such words to only one synset and check the effect it has on the remaining words.

4.8 Summary

In this Chapter we presented our hypothesis for the impact of our contributions to the OWN-EN, the tools chosen to run an experiment to test these hypothesis, their limitations, the results and our remarks on such results. Table 4.4 summarizes how many terms changed synsets within the domains of geological time and lithology as well as outside of such domains.As expected, new words were properly tagged, as shown by *age*, *anhydrite* and *dolostone* numbers. Likewise, as for preexisting words with new senses, results were promising, as in the *Cretaceous* example. The impact on preexisting words for preexisting senses are inconclusive: results ranged from negative like *formation* and *rock* cases, to neutral (*well*, *source* and *basin*) and positive (*Cretaceous* and *sediment*).

	terms	new MWEs
GeoTimes	328	304
Lithology	138	39
Others	375	-

Table 4.4: Synset changes

Chapter 5

Geological time: SUMO extension

Part of this chapter has previously been presented and published[82]. As stated earlier, one of most commonly found information in geoscience material are geological time references, e.g. '165 Million years ago (Ma)', 'during the Jurassic Period,' etc. Applications, such as AgeNames[1] [45] attest the relevance of such information. It was implemented to perform (space and time) query and scan documents for stratigraphic terms, identifying the stratigraphic context of a publication. Should one search for 'Maastrichtian', the application is already able to retrieve documents relating Upper Cretaceous, Cretaceous, Mesozoic and Phanerozoic - respectively the Epoch, Period, Era and Eon containing the Maastrichtian Age. However it does not seem to be able to reason over numeric expressions referring to the interval ranging from 72.1 to 66 million years ago[2], neither to perform even simple inferences. Consider a geoscientist who dated a sample to 70 million year ago: translating this to Maastrichtian Age or retrieving a document that does not mention Maastrichtian but has '70 million years ago' is still not possible within AgeNames or, to the best of our knowledge, other publicly available tools.

Briefly, the reason to use the knowledge representation language SUO-KIF is that a description logic [12] (i.e., the formal language behind OWL [16]) doesn't allow us to capture the original natural language definitions from

[1] http://www.agenames.org/

[2] the time span of the Maastrichtian age

the domain, just the taxonomy of concepts and argument types. OWL does not allow for arities beyond binary, modal statements including temporal qualifications of formulas etc. Without expressive rules supported in SUO-KIF, most of the statements and terms would not be properly formalized, leaving the semantics to the imagination of the user (and each user is likely to have a slightly different intuition), rather than accessible through logical inference.

In this chapter we explain the main concepts of geological time in section 5.1. Section 5.2 shows similar works towards setting ground for time-reasoning as well as the organization, formats for ISC models and variants. Section 5.3 presents our formalization of the domain in SUMO.

5.1 Geological Time

Geological time has its own timescale and it is used by geologists, paleontologists, and other geoscientists to describe the timing and relationships of events in Earth's history. The table of geologic time spans set forth by the International Commission on Stratigraphy (ICS), a sub-committee of the International Union of Geological Sciences, is described in http://www.stratigraphy.org. The geologic timescale is organized in a hierarchical fashion. Eons (or aeons) are divided into eras. Eras contain periods that contain epochs, and finally epochs contain ages. The first three eons (Hadean, Archean, Proterozoic) are collectively referred as the Precambrian super-eon. The most recent eon, the Phanerozoic is subdivided into several periods.

The ICS publishes regularly the International Chronostratigraphic Chart (ICC)[3] as the current standard of the organization of the geologic timescale of the Earth. In the current version, the chart contains 178 names of geological periods. One can read about the development of the chart in [25].

As explained in that paper, geological time periods are not as well-established as one might expect. Most of the systems, series and stages were first defined from type-sections in Europe, the historical home of stratigraphy. Subsequent study of stratigraphical successions worldwide has led to a proliferation of regional units. These historical units did allow Phanerozoic strata to be correlated and mapped worldwide. However, as it happened,

[3]It was previous called International Stratigraphic Chart (ISC). It can be found at http://www.stratigraphy.org/index.php/ics-chart-timescale.

most successive chronostratigraphic units are located in geographically separated type sections, which have more recently been shown to be separated by significant gaps or to overlap considerably. These problems, and the general lack of defined boundaries for historically established units, became serious hindrances to high-resolution correlation of geographically widespread stratigraphic successions.

The ICS committee was tasked with producing a chart that solved the issues of conflicting and overlapping regional strata. We assume the chart and its periods and boundaries represent the current consensus between scientists working on this area, i.e. we assume ICS to be an authority for the concept of geological time. A fragment of the ICC is presented in Figure 5.1.

While the commission was created exactly to unify and organize the classification of both strata and geochronological periods, it appears that the work is not finished and subject to disagreement. In [25] the authors say "[...] disagreement often arises, because type sections that are favored for historical reasons may be abandoned, previously established boundary levels may be greatly changed, and in some instances historical units are replaced by different new ones." Thus while the ontology might look very much a finished product, it seems that its contents are still subject to debate. Another evidence is that between 2012 and 2018 there were eleven different versions of the International Chronostratigraphic Chart.

System/ Period	Series/ Epoch	Stage/ Age	Age (Ma)	
Paleogene	Paleocene	Danian	younger	
Cretaceous	Upper/ Late	**Maastrichtian**	66.0	72.1
		Campanian	72.1	83.6
		Santonian	83.6	86.3
		Coniacian	86.3	89.8
		Turonian	89.8	93.9
		Cenomanian	93.9	100.5
	Lower/ Early	Albian	100.5	~113.0
		Aptian	~113.0	~125.0
		Barremian	~125.0	~129.4
		Hauterivian	~129.4	~132.9
		Valanginian	~132.9	~139.8
		Berriasian	~139.8	~145.0
Jurassic	Upper/ Late	Tithonian	older	

Figure 5.1: A fragment of the ICC presenting the Maastrichtian age.

The geologic timescale is a complex data structure composed of abstract

elements, instants and time intervals, and their relationship with specific concrete representations of geologic records and the observations made of those concrete representations. The International Commission on Stratigraphy recommends a very precise usage of these components in order to establish a standard timescale for global correlations - this has been primarily described in text [83].

The geologic timescale (also geological time) spans from the creation of Earth to Present day. As a convention, *Present* is set as January 1st, 1950. All other points in time are defined according to a set of guidelines and rules adopted by geoscientists and currently under the responsibility of International Commission on Stratigraphy, the largest and oldest constituent scientific body in the International Union of Geological Sciences (IUGS). It aims to define global units (systems, series, and stages) of the International Chronostratigraphic Chart that, in turn, are the basis for the units (periods, epochs, and age) of the International Geologic Time Scale; therefore setting global standards for the fundamental scale for expressing the history of the Earth[4].

In other words, the International Commission on Stratigraphy seeks for markers such as fossils, reversals of the Earth's magnetic field, geochemical signals which can offer a correlation between strata and a point in time. Such markers help define the boundaries between eonothems, erathems, systems, series, and stages. They are the rock strata counterparts of eons, eras, periods, epochs and ages, which are the units of the International Geologic Time Scale [83]. Stratum (plural: strata) is a layer of rock or soil with internally consistent characteristics that make it distinguishable from other layers. A famous example of strata are the layered walls of the Grand Canyon. An example of a marker is a reddish layer at the base of the 50cm thick, dark boundary clay at Oued Djerfane, west of El Kef, Tunisia, which presents an Iridium geochemical anomaly. It marks the precise moment when the asteroid that killed [most] dinosaurs hit Earth. This moment, known as the Cretaceous–Paleogene extinction event, characterizes the end of the Mesozoic Era and the beginning of the Phanerozoic Era [62].

In [28], a representation of the model using the Unified Modeling Language (UML) was presented. The model builds on existing components from standardization of geospatial information systems. From this UML model, [27] presents the ISC ontology, an OWL ontology for the geologic

[4] http://www.stratigraphy.org

timescale derived. All versions of the International Stratigraphic Chart from 2004 to 2014 have been encoded using the ISC ontology. A particular aspect of this ontology is that the elements of the timescale retain the same identifiers across the multiple versions, though the information describing each geochronologic unit evolves with the versions of the timescale. The ISC ontology contains many sub-ontologies including the Geologic Timescale[5] (GTS), the Temporal Hierarchical Ordinal Reference System model[6] (THORS), the Simple Knowledge Organization System (SKOS) [46] and the already mentioned OWL-Time.

Although there are already several approaches to handle ICS Chronostratigraphic Chart, most lack proper support for reasoning. One exception is [76], where two ontologies are developed, one for describing *Geological Time* and another for operating *Geological Dating*, both derived from GeoSciML model [7] to open the possibility of comparing the ages of GeoTimes eventually belonging to different time successions. It, however, does not model sub-periods and is formalized in OWL language, whose limitations were cited early.

Albeit it is possible for one to create a reasoning tool from say, ISC ontology, SUMO already has a backbone for reasoning - it is a formal language with more expressiveness than OWL, with several terms already defined and it includes Allen's Time Interval Algebra [2]. The limitation with ISC ontology is, because it is written in OWL, few logical axioms can be provided beyond the simple taxonomy of concepts. All geological periods are OWL individuals and properties on these instances are defined by "annotation properties". Annotation properties can not be used in property axioms. Thus, in OWL one cannot even define subproperties or domain/range constraints for annotation properties. The object of an annotation property must be either a data literal, a URI reference, or an individual [16]. As we will see in Section 5.3, this imposes a strong limitation in the modeling of the required constraints.

In the GTS ontology, *age, epoch, sub-period, period, era, eon*, and *super-eon* are sub-classes of `GeochronologicEra` (abbreviated as GE). However, there is no formally defined hierarchy between these concepts. Instead, greater emphasis is placed on the boundaries of the periods and, many times,

[5] http://resource.geosciml.org/ontology/timescale/gts.html
[6] http://resource.geosciml.org/ontology/timescale/thors.html
[7] http://www.geosciml.org/

only the approximate duration of the period is given in the chart. It is important to note that geologists qualify the units as "early", "mid", and "late" when referring to time, and "lower", "middle", and "upper" when referring to the corresponding rocks. For example, the lower Jurassic Series in chronostratigraphy corresponds to the early Jurassic Epoch in geochronology. The adjectives are capitalized when the subdivision is formally recognized, and lower case when not; thus "early Miocene" but "Early Jurassic".

The boundaries between periods used to be annotated using the THORS ontology, which is used to define the hierarchy between instances of GE. Fragments of the `ISO19108:2002` standard (Geographic information – temporal schema) are also used to specify the temporal position of geochronologic boundaries[8]. In the more recent versions, THORS ontology properties are mapped to W3C OWL-Time properties. The time interval of a GE is given in terms of its boundaries to other GEs via `time:hasBegin` and `time:hasEnd`. Each boundary is an instance of `gts:GeochronologicBoundary` and it is temporally located via `time:inTemporalPosition` which specifies a `time:numericPosition` with a value, frame (e.g., "Ma"), and a numeric uncertainty when necessary. Nevertheless, the approximate numeric ages provided in the ICS Chart with the (∼) mark were not modeled in the ontology. The boundary modeling should be sufficient for representing the hierarchical relationship between GEs, but ISC ontology further defines an explicit set inclusion relationship between GEs via the `thors:member` property. Also, SKOS is also used to represent inclusion via `skos:narrower`, `skos:broader` along with their transitive versions, `skos:narrowerTransitive` and `skos:broaderTransitive`.

5.2 Temporal Logic

Temporal Logic is a term broadly used to cover all approaches to representing and reasoning about time and temporal information within a logical framework. It can be more narrowly defined to refer the modal-logic introduced by Arthur Prior [79] under the name of Tense Logic and subsequently developed further by many researchers. Over time, Temporal Logic has been used for many applications such as a formalism for clarifying philosophical issues about time, as a framework to precisely define the semantics of temporal

[8] https://www.iso.org/standard/26013.html

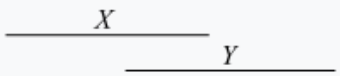
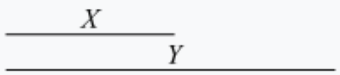
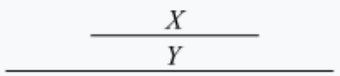
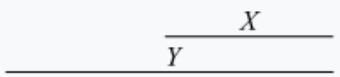
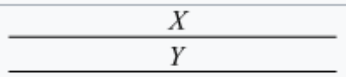
Relation	Illustration	Interpretation
$X < Y$ $Y > X$		X takes place before Y
$X\ \mathbf{m}\ Y$ $Y\ \mathbf{mi}\ X$		X meets Y (*i* stands for *inverse*)
$X\ \mathbf{o}\ Y$ $Y\ \mathbf{oi}\ X$		X overlaps with Y
$X\ \mathbf{s}\ Y$ $Y\ \mathbf{si}\ X$		X starts Y
$X\ \mathbf{d}\ Y$ $Y\ \mathbf{di}\ X$		X during Y
$X\ \mathbf{f}\ Y$ $Y\ \mathbf{fi}\ X$		X finishes Y
$X = Y$		X is equal to Y

Figure 5.2: Allen's Time Interval's Relations [2], figure source: https://en.wikipedia.org/wiki/Allen%27s_interval_algebra, last accessed September 21 2020

expressions in natural language, as a language for encoding temporal knowledge in artificial intelligence and as a tool for specification and verification of computer programs [40].

In a more practical point of view, one of the seminal works is Allen's interval algebra. It is a calculus for temporal reasoning that was introduced in [4]. The calculus defines possible relations between time intervals and provides a composition table that can be used as a basis for reasoning about temporal descriptions of events.

Many vocabularies for time concepts where developed for the Semantic Web initiative. The most notable OWL/RDF vocabulary actively maintained for the time domain is the OWL-Time from W3C[9] but as noted above, lacks the language and reasoning frameworks needed to compute answers to numerical queries about times and dates. Interesting to note is that most of them are derived from the formalization presented in [44], where the presentation is a mix of first order logic formulas and description logics (OWL) formulas and it is not easy to grasp the intended target formal logic language in the paper. For instance, the use of some ternary predicates, such as `timeBetween`, makes the presentation not directly entirely convertible to OWL. The authors say "This effort has been informed by temporal ontologies developed at a number of sites; it is intended to capture the essential features of all of them and make them easily available to a large group of Web developers and users, embedded in the ontology mark-up language OWL.".

5.3 Expanding SUMO with Geochronological Eras

The Suggested Upper Merged Ontology (SUMO) [65] is a formal ontology written in a higher order logic[10]. It is being used for research and applications in search, linguistics and reasoning. It consists of an upper level ontology, a mid-level and dozens of domains ontologies. Together they form roughly 20,000 terms and 80,000 human-authored logical statements. SUMO is the only formal ontology that has been mapped to all of the WordNet lexicon which provides a strong basis for natural language processing applications [66]. There is an associated open source toolset for development,

[9] https://www.w3.org/TR/owl-time/
[10] http://www.ontologyportal.org

debugging and inference on the ontology [74].

SUMO contains most of the content we need for our application, including definitions for time points and intervals and relations between intervals (adapted from [3]). For modeling the geochronological times, we have used two main SUMO classes `TimeInterval` and `TimePoint` and the functions and predicates associated to them.

While a tutorial on the SUO-KIF language is beyond scope here, the interested reader is referred to [73]. In brief, the syntax is valid Lisp S-expressions[11], a prefix notation in which predicates are followed by one or more arguments. Variables are denoted by an initial question mark.

Figure 5.3 presents the definition of the `GeochronologicTime` class and one of its sub-classes, the `GeochronologicSuperEon` class[12]. The remaining sub-classes of `GeochronologicTime` are defined in a similar fashion. Note that all defined classes are sub-classes of the SUMO `TimeInterval` class, inheriting all its properties. Following the definition of the classes we have two important axioms that guarantee the consistency of the model, none of them encoded in the formalization of ISC presented in Section 5.2. The first axiom says that no two instances of `GeochronologicTime` in the same rank can overlap. That is, no two Epoch (Era, Eon, Period etc) can overlap temporally. The second axiom enforces the hierarchical system of time intervals. It says, for instance, that an Age must occur during an Epoch - the remaining axioms for the other classes are similar.

Next, in Figure 5.4, we define the time boundaries between geochronological times. Following the International Commission on Stratigraphy convention, we defined the class `GeochronologicBase` sub-class of the SUMO `TimePoint` class for representing a boundary between periods. The `GeochronologicPresent` constant represents the beginning of the year 1950, taken as the 'current time' by ISC [28]. The function `MillionYearsAgoFn` basically defines the time unit 'Millions of year ago' (Ma). The boundaries between periods can be precisely or approximately defined. In the case of uncertainty, boundaries can be in a range (e.g. 182.7 ± 0.7) or approximations (e.g. 500.5). To represent all these cases we defined three predicates `maBoundary`, `maApproxPoint` and `maPoint` and associated `GeochronologicBase` instances and numbers.

[11] https://en.wikipedia.org/wiki/S-expression

[12] The current full version of Geochronologic Time as an extension of SUMO is found at https://github.com/ontologyportal/sumo/blob/master/GeochronologicTimes.kif

```
(partition GeochronologicTime GeochronologicAge GeochronologicEpoch
           GeochronologicSubPeriod GeochronologicPeriod
           GeochronologicEra
           GeochronologicEon GeochronologicSuperEon)

(subclass GeochronologicTime TimeInterval)
(subclass GeochronologicSuperEon GeochronologicTime)
(termFormat EnglishLanguage GeochronologicSuperEon "supereon")
...

(=>
  (and
    (instance ?X GeochronologicTime)
    (instance ?Y GeochronologicTime)
    (instance ?X ?Class)
    (instance ?Y ?Class)
    (not (equal ?X ?Y))
    (subclass ?class GeochronologicTime))
  (not
    (overlapsTemporally ?X ?Y)))

(=>
  (instance ?X GeochronologicAge)
  (exists (?Y)
    (and
      (instance ?Y GeochronologicEpoch)
      (or
        (starts ?X ?Y)
        (during ?X ?Y)
        (finishes ?X ?Y)))))
...
```

Figure 5.3: GeochronologicTime classes

```
(subclass GeochronologicBase TimePoint)

(instance GeochronologicPresent (BeginFn (YearFn 1950)))

(instance MillionYearsAgoFn UnaryFunction)
(domain MillionYearsAgoFn 1 Number)
(range MillionYearsAgoFn 1 TimePoint)

(equal (MillionYearsAgoFn ?X)
       (BeginFn (YearFn (AdditionFn 1950 (MultiplicationFn ?X -1000000)))))

(instance maBoundary TernaryPredicate)
(domain maBoundary 1 GeochronologicBase)
(domain maBoundary 2 RealNumber)
(domain maBoundary 3 RealNumber)

(=>
  (maBoundary ?Base ?X ?Y)
  (temporallyBetween
    (MillionYearsAgoFn (AdditionFn ?X ?Y))
    ?Base
    (MillionYearsAgoFn (SubtractionFn ?X ?Y))))

(instance maApproxPoint BinaryPredicate)
(domain maApproxPoint 1 GeochronologicBase)
(domain maApproxPoint 2 RealNumber)

(=>
  (maApproxPoint ?Base ?X)
  (exists (?Y)
    (and
      (approximateValue ?X ?Y)
      (equal ?Base (MillionYearsAgoFn ?Y)))))

(instance maPoint BinaryPredicate)
(domain maPoint 1 GeochronologicBase)
(domain maPoint 2 RealNumber)

(=>
  (maPoint ?Base ?X)
  (equal ?Base (MillionYearsAgoFn ?X)))
```

Figure 5.4: GeochronologicTime boundaries

We must emphasize that all predicates used in the previous code fragments, such as `overlapsTemporally`, `during`, `temporallyBetween` etc., are formally defined in SUMO[13] - they are not merely symbols as in the OWL Ontology presented in Section 5.2. Given all the above definitions, we can finally present in Figure 5.5 the SUMO encoding for the fragment of ICS Chart presented in Figure 5.1.

```
1  (instance Maastrichtian GeochronologicAge)
2  (termFormat EnglishLanguage Maastrichtian "Maastrichtian")
3  (termFormat PortugueseLanguage Maastrichtian "Maestrichtiano")
4  (meetsTemporally Campanian Maastrichtian)
5  (meetsTemporally Maastrichtian Danian)
6  (finishes Maastrichtian LateCretaceous)
7  (equal (BeginFn Maastrichtian) BaseMaastrichtian)
8  (equal (EndFn Maastrichtian) BaseCenozoic)
9
10 (instance Danian GeochronologicAge)
11 (termFormat EnglishLanguage Danian "Danian")
12 (termFormat PortugueseLanguage Danian "Daniano")
13 (equal (BeginFn Danian) BaseCenozoic)
14 (equal (EndFn Danian) BaseSelandian)
15
16 (instance BaseMaastrichtian GeochronologicBase)
17 (MaBoundary BaseMaastrichtian 72.1 0.2)
18
19 (instance BaseCenozoic GeochronologicBase)
20 (MaPoint BaseCenozoic 66.0)
```

Figure 5.5: The SUMO encoding of Maastrichtian Age, the SUMO version of the ISC Ontology fragment from Figure 5.1.

It is important to note that Figure 5.1 presents only a small fragment of the axioms added to SUMO. We have expanded SUMO with all the 178 names of geological periods presented in the current version of the International Chronostratigraphic Chart. We used the Maastrichtian example along all steps only to guide the reader.

Given the definitions above, one could employ the SUMO to TFF0 language translation [72][14] available in SigmaKEE [75], with Vampire (or another prover that implements TFF0) to query whether, for example, if 125 Ma is earlier than 113 Ma (as shown in Figure 5.6) or if all the geological periods comply with our axioms. Note that in the proof shown here, the

[13]The definitions can be inspected at http://ontologyportal.org

[14]At the time of this project the translation wasn't working properly. The Author thanks Adam Pease for an ad hoc solution to this issue and for the proof shown here.

type definitions are removed and the proof only shows the axioms from the portion of SUMO needed for the proof. The TFF0 version of SUMO is produced automatically by the Sigma system, and the relevant axioms among the tens of thousands in SUMO are found automatically by Vampire 4.2.2. Axioms marked "axiom" are those from the human-authored SUMO. Axioms marked "plain" are those derived automatically by Vampire. This is a resolution proof, or proof by contradiction, so a successful conclusion is a proof of `$false`. The proof has been simplified to remove trivial steps and allow it to fit on one page.

5.4 Summary

In this chapter we explained the main concepts of geological time, its structure and the authority source chosen for our project - in this case, it is also the authority source for such concept, unlike our choice for lithology. Concerned with setting grounds for time-reasoning tools, we presented similar works and our extension of SUMO onto geological time. It is important to notice that we chose to extend SUMO onto geological time because the limitations of a lexical resource such as a WordNet were so blatant to geological times that we felt obliged to procure and enrich a resource as SUMO with such concepts. We did not identify the same level of limitations for the lithology domain, therefore we chose not to expand SUMO onto lithology and leave it as a future work.

```
tff(f18028,axiom,(! [X0 : $int,X1 : $int,X2,X3] : (($less(X0,X1) &
  equal(X3,s__BeginFn(s__YearFn__1InFn(X1))) &
  equal(X2,s__BeginFn(s__YearFn__1InFn(X0)))) => s__before(X2,X3)))).
tff(f16133,axiom,(! [X0 : $real] : equal(s__MillionYearsAgoFn__1ReFn(X0),
  s__BeginFn(s__YearFn__1InFn(s__FloorFn__0In1ReFn(
    $sum(1950.0,$product(X0,-1000000.0))))))).
tff(f16080,negated_conjecture,(
  ~s__before(s__MillionYearsAgoFn__1ReFn(125.0),
             s__MillionYearsAgoFn__1ReFn(113.0)))).
tff(f16090,axiom,(  ! [X0 : $real] :
s__FloorFn__0In1ReFn(X0) = $to_int(X0))).
tff(f21055,plain,(
  ! [X0 : $int,X1 : $int,X2,X3] : (s__before(X2,X3) | (~$less(X0,X1) |
  ~equal(X3,s__BeginFn(s__YearFn__1InFn(X1))) |
  ~equal(X2,s__BeginFn(s__YearFn__1InFn(X0))))),ennf_trans,[f18028]).
tff(f22979,plain,(  ~s__before(s__MillionYearsAgoFn__1ReFn(125.0),
             s__MillionYearsAgoFn__1ReFn(113.0))),cnf_trans,[f16080]).
tff(f36673,plain,(( ! [X0:$real] :
  (s__FloorFn__0In1ReFn(X0) = $to_int(X0)) )), cnf_trans,[],[f16090]).
tff(f36716,plain,(
  ( ! [X0:$real] : (equal(s__MillionYearsAgoFn__1ReFn(X0),
    s__BeginFn(s__YearFn__1InFn(s__FloorFn__0In1ReFn(
      $sum(1950.0,$product(X0,-1000000.0))))))),cnf_trans,[f16133])))).
tff(f40282,plain,(
  ( ! [X2,X0:$int,X3,X1:$int] : (s__before(X2,X3) | ~$less(X0,X1) |
  ~equal(X3,s__BeginFn(s__YearFn__1InFn(X1))) |
  ~equal(X2,s__BeginFn(s__YearFn__1InFn(X0)))) )),cnf_trans,[f21055]).
tff(f40348,plain,(
  ( ! [X0:$real] : (equal(s__MillionYearsAgoFn__1ReFn(X0),
    s__BeginFn(s__YearFn__1InFn($to_int(
      $sum(1950.0,$product(X0,-1000000.0)))))) )),
  definition_unfolding,[f36716,f36673]).
tff(f40413,plain,(
  ( ! [X4:$int,X5:$int] : (~equal(s__MillionYearsAgoFn__1ReFn(113.0),
    s__BeginFn(s__YearFn__1InFn(X5))) | ~$less(X4,X5) |
    ~equal(s__MillionYearsAgoFn__1ReFn(125.0),
           s__BeginFn(s__YearFn__1InFn(X4)))) )),
  resolution,[f22979,f40282]).
tff(f40594,plain,(  ( ! [X0:$int] : (~$less(X0,$to_int(
    $sum(1950.0,$product(113.0,-1000000.0)))) |
    ~equal(s__MillionYearsAgoFn__1ReFn(125.0),
           s__BeginFn(s__YearFn__1InFn(X0)))) )),
  resolution,[f40413,f40348]).
tff(f40664,plain,(( ! [X0:$int] : (~equal(s__MillionYearsAgoFn__1ReFn(125.0),
          s__BeginFn(s__YearFn__1InFn(X0))) | ~$less(X0,-112998050)) )),
  evaluation,[f40594]).
tff(f40665,plain,(
  ~$less($to_int($sum(1950.0,$product(125.0,-1000000.0))),-112998050)),
  resolution,[f40664,f40348]).
tff(f40734,plain,(~$less($to_int(-124998050.0),-112998050)),
  evaluation,[f40665]).
tff(f40735,plain,($false),evaluation,[f40734]).
```

Figure 5.6: A Simplified Proof in the TFF0 Version of SUMO with the Vampire Prover as shown in [82]

Chapter 6

Conclusion

This project set out to contribute to NLP applied to Oil&Gas domain, particularly to the petroleum geology realm. While there are many works towards advancing algorithms for each and every aspect of the NLP "pipeline", some even to the pipeline as a whole, we chose to focus into the resources that support such algorithms. In particular, we chose to focus on expanding the lexical resource known as WordNet. In Chapter 2 we described the original Princeton WordNet as well as several initiatives to extend it both vertically, into specific domains, and horizontally, onto both other languages and English, since Princeton WordNet development has halted.

In Chapter 3 we vertically extended OWN-EN and OWN-PT with terms related to lithology and geological time, the two main concepts of geosciences. In Chapter 5 we also extended SUMO with the geological time concepts, hoping to set ground for the eventual development of tools which could reason over time. All such contributions are freely available online at https://github.com/own-pt/own-en, https://github.com/own-pt/own-pt and https://github.com/ontologyportal/sumo/. It is worth noticing we kept the links to other resources that were originally within GBA material, believing that this will enable future users to build not only upon WordNet but also upon these other valuable material - the mapping file linking OWN-EN synset ids (which are the same of PWN 3.0 were applicable), OWN-EN sense keys, GBA Thesaurus URI, BGS, INSPIRE, geosciml, DBPedia and wikidata is available at https://github.com/alexandretessarollo/MSc, the same URL for our corpus in both its annotated and raw versions.

Another contribution of this project was the first large scale test of both

the textual approach to WordNet and the mill tool[1]. Throughout this undertaking some issues were detected and signaled to mill, as were some improvement suggestions. As for the textual-file-approach to WordNet editing, our first-hand experience allows us to say it is a successful approach. The data structure is intuitive and easy to work with, and one can choose to work it both manually with text processors or programmatically with scripts in Python or any other computer language capable of processing string files. OWN-EN in particular has a sensemap to Princeton WordNet 3.0 synset ids[2], which allowed the exercise we conducted in our Evaluation - one can read "PWN 3.0" instead of "original OWN-EN" throughout chapter 4. However, the textual files were constructed with a synset-id-less approach, which complicated the evaluation to a point where we were not able to use the WNDB files generated by mill. One suggestion of improvement to the OWN-EN and OWN-PT projects is to adopt some sort of unique synset-id, preferably one that links to PWN 3.0 where possible. Our suggestion is to set it as another relation such as "id: 14696793-n".

Chapter 4 confirmed our basic expectations regarding both new words and new senses to preexisting words within OWN-EN. It also showed promising results regarding our hypothesis on preexisting words for preexisting senses, i.e., the ones that weren't affected in OWN-EN by our contributions.

Relating to our SUMO extension, and considering geological time is subdivided in intricate ways and its modeling is a work in progress, we believe this work can contribute to updates and improvements of the ISC ontology. With our SUMO extension we were able to clarify some points in the most 2018 published version of ISC Ontology such as Capitanian Age and Upper Mississippian Sub Period inconsistent endings and the missing information about the approximate numeric ages. One can now use SUMO as a geological time enriched resource to develop tools that use time reasoning for the Oil&Gas domain.

6.1 Reproducibility

In order to reproduce our work in Chapter 3, whether to the domains we chose or others, one must choose an authoritative source, keeping in mind the *authority* comes from the choice itself. It should be a source with no licensing

[1] https://github.com/own-pt/mill

[2] https://github.com/own-pt/own-en/blob/master/etc/sensemap.tsv

limitations, if not complete with at least a comprehensive coverage of terms and definitions, and already structured in an ontology-like format - or at least "structurable", where one would need to infer the relations somehow, like we did with the meronym relations from reading of definitions. The next step is to map the relations in the chosen source with those of the chosen WordNet project to extend. Then one must evaluate on where to append the new terms[3]. Finally comes the merging step, where new definitions will be easy to include, but already existing definitions will pose the challenge of being identified[4] and having the appropriateness of their definition and relations assessed.

These same macro steps apply to extending other resources like SUMO, for instance. Other than the authoritative source and the lexical resource to extend, one last key component is having human resources who have a reasonable degree of knowledge of both the resource being extended as of the chosen domain.

Replicating our extensions to other languages is, as explained in Section 3.5, rather trivial for geological time: the terms have official translations and the definitions used follow a simple frame which should be simple to translate. If necessary, one may suppress the "distinguished by" part of the definition to ease the process. However, for the lithological concepts it might prove challenging: the definitions from GBA were adjusted to fit WordNet, and the use of INSPIRE for the translation into Portuguese showed that most of definitions were not translated and the few ones translated also needed adjustments. So, translating the definitions will require the aid of a bilingual Oil&Gas expert. As for translating the terms, INSPIRE does translate them to European languages, but at least in Portuguese it is not a translation that abide to geological naming rules. Because GBA is both in English and German and was built by Austrians, one might expect it to have German terms conforming to such naming rules.

[3]With lithology, for instance, we chose to add the materials and then add all other terms connecting them to materials and eventually to rock

[4]some will be like the *sand* example, associated with a term in the chosen source, other will be like *arenaceous rock*, absent from the source but with a definition that requires checking

6.2 Future Works

A future work is confirming our results of Chapter 4. Some of the ways envisioned to do so involve testing the current OWN-EN over different corpora and not only concluding the Princeton WordNet Gloss Corpus Project[5], but also extending it by annotating the very definitions we included or altered in this project.

A similar experiment could be done in Portuguese. One relevant set of corpora to be used can be found at http://petroles.ica.ele.puc-rio.br [26].

Among other future works is producing a golden set[6] over Oil&Gas data sets with the current OWN-EN and running the evaluation again on it to assess WSD choices over previously existing words. Improving the UKB algorithm implemented in Freeling to consider the "one sense per discourse"[38] / "one sense per collocation [MWE]"[96] aspect is another future work which may improve the WSD results.

The English WordNet [53] has larger reach as a WordNet in English than OWN-EN. Transferring the contributions made to OWN-EN onto English WordNet will be a major milestone in our project.

The presented SUMO encoding of geological time opens possibility of future works on the formalization of other important domain specific information artifacts, such as a chronostratigraphic chart of a given area. Also as future work, there is still the need to encode in SUMO the stratotype or type sections. Stratotypes are physical locations or outcrop of a particular reference exposure of a stratigraphic sequence or stratigraphic boundary; they are represented in the ISC ontology.

Within geological time, we are currently in touch with leading researchers in ICS-related studies, with ongoing discussions to consolidate one single point of truth for timescale and GSSP information from which PDF charts, the stratigraphy.org website etc could ultimately all be generated.

During our SUMO expansion we came to realize the TFF0 language translation available in SigmaKEE was not working properly. Adam Pease, the Technical Editor of SUMO, provided instrumental assistance to our specific necessities in this project, but the TFF0 issue is still open and solving it is a future work.

[5] a corpus of manually annotated synset definitions (glosses) from the PWN, available at http://wordnetcode.princeton.edu/glosstag.shtml

[6] a corpus annotated with proper senses

References

[1] Eneko Agirre and Aitor Soroa. "Personalizing PageRank for Word Sense Disambiguation". In: *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*. EACL '09. Athens, Greece: Association for Computational Linguistics, Jan. 2009, pp. 33–41. DOI: 10.3115/1609067.1609070. URL: https://dl.acm.org/doi/10.5555/1609067.1609070.

[2] James Allen. *Natural language understanding*. The Benjamin/Cummings Publishing Company, 1995. ISBN: 0805303340.

[3] James Allen. "Towards a general theory of action and time". In: *Artificial Intelligence* 23 (1984), pp. 123–154.

[4] James F. Allen. "Maintaining Knowledge About Temporal Intervals". In: *Commun. ACM* 26.11 (Nov. 1983), pp. 832–843. ISSN: 0001-0782. DOI: 10.1145/182.358434. URL: http://doi.acm.org/10.1145/182.358434.

[5] Raquel Amaro and Sara Mendes. "Towards merging common and technical lexicon wordnets". In: *Proceedings of the 3rd Workshop on Cognitive Aspects of the Lexicon*. Mumbai, India: The COLING 2012 Organizing Committee, Dec. 2012, pp. 147–160. URL: https://www.aclweb.org/anthology/W12-5112.

[6] M Antoniak, J Dalgliesh, M Verkruyse, and J Lo. "Natural Language Processing Techniques on Oil and Gas Drilling Data". In: *Intelligent Energy International Conference*. Sept. 2016, pp. 1–6.

[7] M Antoniak, J Dalgliesh, M Verkruyse, J Lo, et al. "Natural Language Processing Techniques on Oil and Gas Drilling Data". In: *SPE Intelligent Energy International Conference and Exhibition*. Society of Petroleum Engineers. 2016.

[8] Mihael Arcan, Christian Federmann, and Paul Buitelaar. "Using Domain specific and Collaborative Resources for Term Translation". In: *Proceedings of the Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation*. Jeju, Republic of Korea: Association for Computational Linguistics, July 2012, pp. 86–94. URL: https://www.aclweb.org/anthology/W12-4210.

[9] Robert Arp, Barry Smith, and Andrew D. Spear. *Building Ontologies with Basic Formal Ontology*. The MIT Press, 2015. ISBN: 0262527812.

[10] Ro Artale, Anna Goy, Bernardo Magnini, and Emanuele Pianta. "Coping with wordnet sense proliferation". In: *In First International Conference on Language Resources & Evaluation*. 1998.

[11] MFJ van Assem, Aldo Gangemi, and AT Schreiber. *RDF/OWL representation of WordNet*. Tech. rep. https://www.w3.org/TR/wordnet-rdf/. World-Wide Web Consortium W3C, 2006.

[12] Franz Baader, Diego Calvanese, Deborah McGuinness, Daniele Nardi, and Peter F. Patel-Schneider. *The Description Logic Handbook: Theory, Implementation and Applications*. 2010. ISBN: 978-0521150118.

[13] Satanjeev Banerjee and Ted Pedersen. "An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet". In: *Computational Linguistics and Intelligent Text Processing*. Ed. by Alexander Gelbukh. Berlin, Heidelberg: Springer Berlin Heidelberg, 2002, pp. 136–145. ISBN: 978-3-540-45715-2.

[14] Anabela Barreiro and Fernando Batista. "Machine Translation of Non-Contiguous Multiword Units". In: *Proceedings of the Workshop on Discontinuous Structures in Natural Language Processing*. San Diego, California: Association for Computational Linguistics, June 2016, pp. 22–30. DOI: 10.18653/v1/W16-0903. URL: https://www.aclweb.org/anthology/W16-0903.

[15] Ander Barrena, Eneko Agirre, Bernardo Cabaleiro, Anselmo Peñas, and Aitor Soroa. ""One Entity per Discourse" and "One Entity per Collocation" Improve Named-Entity Disambiguation". In: *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. Dublin, Ireland: Dublin City University and Association for Computational Linguistics, Aug. 2014, pp. 2260–2269. URL: https://www.aclweb.org/anthology/C14-1213.

[16] Sean Bechhofer, Frank van Harmelen, Jim Hendler, Ian Horrocks, Deborah L. McGuinness, Peter F. Patel-Schneider, and Lynn Andrea Stein. 2004. URL: https://www.w3.org/TR/owl-ref/.

[17] Luisa Bentivogli, Pamela Forner, Bernardo Magnini, and Emanuele Pianta. "Revising WordNet Domains Hierarchy: Semantics, Coverage, and Balancing". In: *Proceedings of COLING 2004 Workshop on "Multilingual Linguistic Resources"*. 2004, pp. 101–108.

[18] Francis Bond and Arthur Bond. "GeoNames Wordnet (gnwn): extracting wordnets from GeoNames". In: *Wordnet Conference*. 2019, p. 387.

[19] Francis Bond and Ryan Foster. "Linking and Extending an Open Multilingual Wordnet". In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Sofia, Bulgaria: Association for Computational Linguistics, Aug. 2013, pp. 1352–1362. URL: https://www.aclweb.org/anthology/P13-1133.

[20] Francis Bond and Kyonghee Paik. "A Survey of WordNets and their Licenses". In: *Proceedings of the 6th Global WordNet Conference (GWC 2012)*. 64–71. Matsue, 2012. URL: https://www.researchgate.net/publication/267427763_A_Survey_of_WordNets_and_their_Licenses.

[21] Francis Bond, Piek Vossen, John P McCrae, and Christiane Fellbaum. "Cili: the collaborative interlingual index". In: *Proceedings of the 8th Global WordNet Conference 2016*. Vol. 2016. 2016. URL: https://www.overleaf.com/read/rsnvsbdghybg.

[22] Sergey Brin and Lawrence Page. "The Anatomy of a Large-Scale Hypertextual Web Search Engine". In: *Proceedings of the Seventh International Conference on World Wide Web 7*. WWW7. Brisbane, Australia: Elsevier Science Publishers B. V., 1998, pp. 107–117.

[23] Paul Buitelaar and Bogdan Sacaleanu. "Extending Synsets with Medical Terms". In: *Proceedings of the First International Conference on Global WordNet* (2002), pp. 21–25.

[24] Yoonjung Choi, Youngho Kim, and Sung-Hyon Myaeng. "Domain-Specific Sentiment Analysis Using Contextual Feature Generation". In: *Proceedings of the 1st International CIKM Workshop on Topic-Sentiment Analysis for Mass Opinion*. TSA '09. Hong Kong, China:

Association for Computing Machinery, 2009, pp. 37–44. DOI: 10.1145/1651461.1651469. URL: https://doi.org/10.1145/1651461.1651469.

[25] Kim Cohen, Stan Finney, Phil Gibbard, and Junxuan Fan. "The ICS International Chronostratigraphic Chart". In: *Episodes* 36.3 (2013), pp. 199–204.

[26] Fábio Cordeiro. *Petrolês - Como Construir um Corpus Especializado em Óleo e Gás em Português*. PUC-Rio, Rio de Janeiro, RJ - Brasil: PUC-Rio, 2020.

[27] S J D Cox and S M Richard. "A geologic timescale ontology and service". In: *Earth Science Informatics* 8.1 (July 2014), pp. 5–19.

[28] Simon J D Cox and Stephen M Richard. "A formal model for the geologic time scale and global stratotype section and point, compatible with geospatial information transfer standards". In: *Geosphere* 1.3 (2005), pp. 119–19.

[29] Daniele Di Giorgio. *Fatores geológicos no planejamento de lavra de rochas ornamentais*. 2003. URL: http://hdl.handle.net/10183/3045.

[30] American Heritage Dictionary. *The American Heritage Science Dictionary*. Houghton Mifflin Company, 2005. ISBN: 9780618455041. URL: https://books.google.com.br/books?id=yKUagx8PB%5C_EC.

[31] The Editors of Encyclopaedia Britannica. *Limestone*. Encyclopædia Britannica, inc., 2018. URL: https://www.britannica.com/science/limestone. (accessed: 12.01.2020).

[32] J. Feblowitz. "Analytics in Oil and Gas: The Big Deal About Big Data". In: *SPE Digital Energy Conference and Exhibition*. The Woodlands, Texas, Mar. 2013.

[33] Christiane Fellbaum. "English Verbs as a Semantic Net". In: *International Journal of Lexicography* 3 (Jan. 1993). DOI: 10.1093/ijl/3.4.278.

[34] Christiane Fellbaum, ed. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press, 1998. ISBN: 026206197X.

[35] Olivier Ferret. "Building a network of topical relations from a corpus". In: *LREC*. 2006, pp. 575–580. URL: https://www.cs.brandeis.edu/~marc/misc/proceedings/lrec-2006/pdf/500_pdf.pdf.

[36] W N. Francis and H. Kučera. "Brown Corpus Manual". In: (1982). URL: http://icame.uib.no/brown/bcm.html.

[37] W. N. Francis and H. Kučera. "Frequency analysis of English Usage: Lexicon and Grammar". In: (1982).

[38] William A. Gale, Kenneth W. Church, and David Yarowsky. "One Sense per Discourse". In: *Proceedings of the Workshop on Speech and Natural Language*. HLT '91. Harriman, New York: Association for Computational Linguistics, 1992, pp. 233–237. ISBN: 1558602720. DOI: 10.3115/1075527.1075579. URL: https://doi.org/10.3115/1075527.1075579.

[39] Spandana Gella, Carlo Strapparava, and Vivi Nastase. "Mapping WordNet Domains, WordNet Topics and Wikipedia Categories to Generate Multilingual Domain Specific Resources". In: *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*. 2014, pp. 1117–1121.

[40] Valentin Goranko and Antony Galton. "Temporal Logic". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Winter 2015. Metaphysics Research Lab, Stanford University, 2015.

[41] Derek Gross and Katherine Miller. "Adjectives in WordNet". In: *International Journal of Lexicography* 3 (Dec. 1990). DOI: 10.1093/ijl/3.4.265.

[42] Reem Al-Halimi, Robert C. Berwick, J. F. M. Burg, Martin Chodorow, Christiane Fellbaum, Joachim Grabowski, Sanda Harabagiu, Marti A. Hearst, Graeme Hirst, Douglas A. Jones, Rick Kazman, Karen T. Kohl, Shari Landes, Claudia Leacock, George A. Miller, Katherine J. Miller, Dan Moldovan, Naoyuki Nomura, Uta Priss, Philip Resnik, David St-Onge, Randee Tengi, Reind P. van de Riet, and Ellen Voorhees. *WordNet An Electronic Lexical Database*. Ed. by Christiane Fellbaum. 1998. ISBN: 9780262061971.

[43] Sanda M. Harabagiu, George A. Miller, and Dan I. Moldovan. "Wordnet 2: a morphologically and semantically enhanced resource". In: *Proceedings of SIGLEX99: Standardizing Lexical Resources*. 1999, pp. 1–8.

[44] Jerry R Hobbs and Feng Pan. "An Ontology of Time for the Semantic Web". In: *ACM Transactions on Asian Language Information Processing* 3.1 (Mar. 2004), pp. 66–85.

[45] Robert Huber and Jens Klump. "Agenames a stratigraphic information harvester and text parser". In: *Earth Science Informatics* 8 (1 2015), pp. 125–134. URL: https://doi.org/10.1007/s12145-014-0171-5.

[46] Antoine Isaac and Ed Summers. *SKOS Primer*. Tech. rep. latest version available at http://www.w3.org/TR/skos-primer. W3C, 2008.

[47] Donatien Ishimwe. *Reservoir rock properties*. URL: https://connect.spe.org/blogs/donatien-ishimwe/2014/09/29/reservoir-rock-properties.

[48] Dan Jurafsky and James H. Martin. *Speech and Language Processing*. 3rd draft. Draft, 2019. URL: https://web.stanford.edu/~jurafsky/slp3/.

[49] H. Kučera and W. N. Francis. "Computational analysis of present-day American English". In: (1967).

[50] Michael Lesk. "Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone". In: *Proceedings of the 5th Annual International Conference on Systems Documentation*. SIGDOC '86. Toronto, Ontario, Canada: ACM, 1986, pp. 24–26. ISBN: 0-89791-224-1. DOI: 10.1145/318723.318728. URL: http://doi.acm.org/10.1145/318723.318728.

[51] Russell J. Lundholm, Rafael Rogo, and Jenny Li Zhang. "Restoring the Tower of Babel: How Foreign Firms Communicate with U.S. Investors". In: *The Accounting Review* 89.4 (2014), pp. 1453–1485. DOI: 10.2308/accr-50725. eprint: https://doi.org/10.2308/accr-50725. URL: https://doi.org/10.2308/accr-50725.

[52] Bernardo Magnini and Gabriela Cavaglia. "Integrating Subject Field Codes into WordNet". In: *Proceedings of LREC-2000*. 2000, pp. 1413–1418. URL: http://wndomains.fbk.eu/publications/lrec-2000.pdf.

[53] John P. McCrae, Alexandre Rademaker, Francis Bond, Ewa Rudnicka, and Christiane Fellbaum. "English WordNet 2019 – An Open-Source WordNet for English". In: *Proceedings of the 10th Global WordNet Conference GWC 2019*. Global Wordnet Association. July 23, 2019. URL: http://john.mccr.ae/papers/mccrae2019english.pdf.

[54] John P. McCrae, Ian Wood, and Amanda Hicks. "The Colloquial WordNet: Extending Princeton WordNet with Neologisms". In: *Proceedings of the First Conference on Language, Data and Knowledge (LDK2017)*. 2017, pp. 194–202. DOI: 10.1007/978-3-319-59888-8_17.

[55] John Philip McCrae, Alexandre Rademaker, Ewa Rudnicka, and Francis Bond. "English WordNet 2020: Improving and Extending a WordNet for English using an Open-Source Methodology". In: *Proceedings of the Multimodal Wordnets Workshop at LREC 2020*. May 11, 2020.

[56] George A. Miller. "Session Summary". In: *Proceedings of the Workshop on Human Language Technology*. HLT '94. Plainsboro, NJ: Association for Computational Linguistics, 1994, p. 7. ISBN: 1558603573. DOI: 10.3115/1075812.1075815. URL: https://doi.org/10.3115/1075812.1075815.

[57] George A. Miller. "The association of ideas". In: *The General Psychologist* 29 (1993), pp. 69–74.

[58] George A Miller, Claudia Leacock, Randee Tengi, and Ross T Bunker. "A semantic concordance". In: *Proceedings of the workshop on Human Language Technology*. Association for Computational Linguistics. 1993, pp. 303–308. URL: https://dl.acm.org/doi/10.3115/1075671.1075742.

[59] George Miller, R. Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller. "Introduction to WordNet: An On-line Lexical Database*". In: 3 (Jan. 1991). DOI: 10.1093/ijl/3.4.235.

[60] George Miller and Florentina Hristea. "WordNet Nouns: Classes and Instances". In: *Computational Linguistics* 32 (Mar. 2006), pp. 1–3. DOI: 10.1162/coli.2006.32.1.1.

[61] Dan Moldovan and Adrian Novischi. "Word sense disambiguation of WordNet glosses". In: *Computer Speech & Language* 18.3 (2004), pp. 301–317.

[62] Eustoquio Molina, Laia Alegret, Ignacio Arenillas, Jose A. Arz, Njoud Gallala, Jan Hardenbol, Katharina von Salis, Etienne Steurbaut, Noel Vandenberghe, and Dalila Zaghbib-Turki. "The Global Boundary Stratotype Section and Point for the base of the Danian Stage (Paleocene, Paleogene, Tertiary, Cenozoic) at El Kef, Tunisia — Original definition and revision". In: *Episodes* 29.4 (Dec. 2006), pp. 263–273. DOI: 10.18814/epiiugs/2006/v29i4/004. URL: https://doi.org/10.18814/epiiugs/2006/v29i4/004.

[63] Henrique Muniz, Fabricio Chalub, Alexandre Rademaker, and Valeria de Paiva. "Extending Wordnet to Geological Times". In: *Global Wordnet Conference 2018*. Singapore, Jan. 2018.

[64] Isaac Newton. "Letter from Sir Isaac Newton to Robert Hooke". In: *Historical Society of Pennsylvania* (1676).

[65] I. Niles and A. Pease. "Toward a Standard Upper Ontology". In: *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems*. Ed. by Chris Welty and Barry Smith. FOIS-2001, 2001.

[66] Ian Niles and Adam Pease. "Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology". In: *Proceedings of the IEEE International Conference on Information and Knowledge Engineering*. 2003, pp. 412–416.

[67] Lluis Padro and Evgeny Stanilovsky. "Freeling 3.0: Towards wider multilinguality". In: *LREC2012*. 2012.

[68] Valeria de Paiva, Alexandre Rademaker, and Gerard de Melo. "Open WordNet-PT: An Open Brazilian Wordnet for Reasoning". In: *Proceedings of COLING 2012: Demonstration Papers*. Published also as Tech report http://hdl.handle.net/10438/10274. Mumbai, India: The COLING 2012 Organizing Committee, Dec. 2012, pp. 353–360. URL: http://www.aclweb.org/anthology/C12-3044.

[69] Betsy Palkowsky. "A New Approach to Information Discovery". In: *SPE Annual Technical Conference and Exhibition*. Dallas, Texas: Society of Petroleum Engineers, Oct. 2005.

[70] Betsy Palkowsky. "A New Approach to Information Discovery - Geography Really Does Matter". In: *SPE Annual Technical Conference*. Dallas, Oct. 2005, pp. 9–12.

[71] European Parliament and of the Council. "INSPIRE Directive 2007/2/EC". In: *Official Journal of the European Union*. Vol. 50. 2007, 371–es. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L:2007:108:TOC.

[72] Adam Pease. "Arithmetic and Inference in a Large Theory (extended abstract)". In: *4th Conference on Artificial Intelligence and Theorem Proving*. 2019.

[73] Adam Pease. *Ontology: A Practical Guide*. Angwin, CA: Articulate Software Press, 2011.

[74] Adam Pease and Christoph Benzmuller. "Sigma: An Integrated Development Environment for Logical Theories". In: *AI Communications* 26 (2013), pp. 9–97.

[75] Adam Pease and Stephan Schulz. "Knowledge Engineering for Large Ontologies with Sigma KEE 3.0". In: *The International Joint Conference on Automated Reasoning*. 2014.

[76] Michel Perrin, Laura S Mastella, Olivier Morel, and Alexandre Lorenzatti. "Geological time formalization: an improved formal model for describing time successions and their correlation". In: *Earth Science Informatics* 4.2 (2011), pp. 81–96.

[77] Wim Peters, Maria Teresa Sagri, Daniela Tiscornia, and Sara Castagnoli. "The LOIS project". In: *LREC*. Jan. 2006. URL: http://www.lrec-conf.org/proceedings/lrec2006/pdf/582_pdf.pdf.

[78] Michael Poprat, Elena Beisswanger, and Udo Hahn. "Building a BioWordNet by Using WordNet's Data Formats and WordNet's Software Infrastructure: A Failure Story". In: *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*. SETQA-NLP '08. Columbus, Ohio: Association for Computational Linguistics, 2008, pp. 31–39. ISBN: 9781932432107.

[79] Arthur Prior. "Tense Logic and the Continuity of Time". In: *Studia Logica* 13 (1962), pp. 133–148.

[80] Alexandre Rademaker. "Challenges for Information Extraction in the Oil and Gas Domain". In: *Proceedings of the XI Seminar on Ontology Research in Brazil (ONTOBRAS)*. Ed. by Joel Luís Carbonera, Giancarlo Guizzardi, Sandro Rama Fiorini, and Mara Abel. São Paulo, Brazil, 2018. URL: http://ceur-ws.org/Vol-2228/invited1.pdf.

[81] Alexandre Rademaker, Bruno Cuconato, Henrique Muniz, Alexandre Tessarollo, and Alessandra Cid. "Completing the Princeton Annotated Gloss Corpus Project". In: *Wordnet Conference*. CLARIN-PL digital repository. Global Wordnet Association. July 2019, p. 378. URL: http://hdl.handle.net/11321/718.

[82] Alexandre Rademaker, Alexandre Tessarollo, and Henrique Muniz Adam Pease. "Extending SUMO to Geological Times". In: *Proceedings of the XII Seminar on Ontology Research in Brazil*. Ed. by João Paulo A. Almeida, Marcello Bax, Rita Berardi, and Fernanda Baião. Vol. 2519. See http://ceur-ws.org/Vol-2519/. Porto Alegre, RS, Sept. 2019, pp. 70–82.

[83] Jurgen Remane, Michael G Bassett, John Cowie, Klaus H Gohrbandt, H Richard Lane, Olaf Michelsen, and Wang Naiwen. "Revised guidelines for the establishment of global chronostratigraphic standards by the International Commission on Stratigraphy (ICS)". In: *Episodes* 19.3 (1996), pp. 77–81.

[84] Adriana Roventini, Alone Antonietta, Francesca Bertagna, Nicoletta Calzolari, Cacila Jessica, Christian Girardi, Bernardo Magnini, R Marinelli, Manuela Speranza, and A Zampolli. "Italwordnet: building a large semantic database for the automatic treatment of Italian". In: *Linguistica computazionale : XVIII/XIX, 1998/1999* (2003). URL: https://doi.org/10.1400/18178.

[85] Maria Teresa Sagri, Daniela Tiscornia, and Francesca Bertagna. *JurWordNet*. 2004.

[86] Eric Schoen, Reid Smith, and John Boden. "AI Supports Information Discovery and Analysis in an SPE Research Portal". In: *SPE Annual Technical Conference and Exhibition*. Society of Petroleum Engineers, Sept. 2018. DOI: 10.2118/191758-MS. URL: https://doi.org/10.2118/191758-MS.

[87] Barry Smith and Christiane Fellbaum. "Medical WordNet: A New Methodology for the Construction and Validation of Information Resources for Consumer Health". In: *Proceedings of the 20th International Conference on Computational Linguistics*. COLING '04. Geneva, Switzerland: Association for Computational Linguistics, 2004, 371–es. DOI: 10.3115/1220355.1220409. URL: https://doi.org/10.3115/1220355.1220409.

[88] Denis Roberto de Souza et al. *Distribuição dos minerais pesados ao longo do curso inferior do rio Itajaí-Açu/SC e sua correlação sedimentar com a planície costeira, praia e plataforma continental interna adjacente*. Florianópolis, SC, 2007. URL: https://repositorio.ufsc.br/bitstream/handle/123456789/89877/245563.pdf.

[89] Randee Tengi. "Design and Implementation of the WordNet Lexical Database and Searching Software". In: Jan. 1998, pp. 105–127.

[90] Alexandre Tessarollo and Alexandre Rademaker. "Inclusion of Lithological terms (rocks and minerals) in The Open Wordnet for English". In: *Proceedings of the LREC 2020 Workshop on Multimodal Wordnets (MMW2020)*. Ed. by Thierry Declerck, Itziar Gonzalez-Dios, and German Rigau. Marseille, France: The European Language Resources Association (ELRA), May 2020, pp. 33–38. ISBN: 979-10-95546-41-2. URL: https://www.aclweb.org/anthology/2020.mmw-1.6.

[91] Alexandre Tessarollo and Alexandre Rademaker. "Information extraction on financial reports of oil and gas industry,: First steps". In: *AAAI 2019 Spring Symposium: Towards AI for Collaborative Open Science*. Stanford University, USA, Mar. 2019. URL: https://aaai.org/Symposia/Spring/sss19symposia.php#ss07.

[92] Zygmunt Vetulani and Bartlomiej Kochanowski. ""PolNet - Polish WordNet" project: PolNet 2.0 - a short description of the release". In: *Proceedings of the Seventh Global Wordnet Conference*. Tartu, Estonia: University of Tartu Press, Jan. 2014, pp. 400–404. URL: https://www.aclweb.org/anthology/W14-0155.

[93] Zygmunt Vetulani, Justyna Walkowska, Tomasz Obrebski, Jacek Marciniak, Pawel Konieczka, and Przemyslaw Rzepecki. "An Algorithm for Building Lexical Semantic Network and Its Application to PolNet - Polish WordNet Project". In: *Human Language Technology. Challenges of the Information Society*. Ed. by Zygmunt Vetulani and Hans Uszkoreit. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 369–381. ISBN: 978-3-642-04235-5. URL: https://doi.org/10.1007/978-3-642-04235-5_32.

[94] Piek Vossen. *EuroWordNet General Document. EuroWordNet (LE2-4003, LE4-8328), Part A, Final Document*. 2002. URL: https://research.vu.nl/ws/files/77020259/EWNGeneral.

[95] PJTM Vossen. "EuroWordNet: a multilingual database for information retrieval". In: *Proceedings of the DELOS workshop on Cross-language Information Retrieval, March 5-7, 1997 Zurich*. Vrije Universiteit. 1997. URL: https://research.vu.nl/ws/files/73708632/Delos97.

[96] David Yarowsky. "One Sense per Collocation". In: *Proceedings of the Workshop on Human Language Technology*. HLT '93. Princeton, New Jersey: Association for Computational Linguistics, 1993, pp. 266–271. ISBN: 1558603247. DOI: 10.3115/1075671.1075731. URL: https://doi.org/10.3115/1075671.1075731.