An adaptation of Text2Onto for supporting the French language

Morad Hajji, Mohammed Qbadou, Khalifa Mansouri
Signals, Distributed Systems and Artificial Intelligence laboratory (SSDIA), University Hassan II Casablanca, Morocco

Article Info	ABSTRACT
	The ontologies are progressively imposing themselves in the field of knowledge management. While the manual construction of an ontology is by far the most reliable, this task has proved to be too tedious and expensive. To assist humans in the process of building an ontology, several tools have emerged proposing the automatic or semi-automatic construction of ontologies. In this context, Text2Onto has become one of the most recognized ontology learning tools. The performance of this tool is confirmed by several research works. However, the development of this tool is based on Princeton WordNet (PWN) for English. As a result, it is limited to the processing of textual resources written in English. In this paper, we present our approach based on JWOLF, a Java API to access the free WordNet for French that we have developed to adapt this tool for the construction of ontologies from corpus in French. To evaluate the usefulness of our approach, we assessed the performance of the improved version of Text2Onto on a simplistic corpus of French language documents. The results of this experiment have shown that the improved version of Text2Onto according to our approach is effective for the construction of an ontology from textual documents in the French language.

Corresponding Author:

Hajji Morad,
Signals, Distributed Systems and Artificial Intelligence laboratory (SSDIA),
University Hassan II Casablanca,
Morocco.
Email: morad.hajji@gmail.com

1. INTRODUCTION

Historically, the origins of the term "ontology" date back to the time of philosophers Plato and Aristotle (400-360 BC). This term is composed of the Greek linguistic roots "ontos" which means to be and "logos" which means science. Thus, "Ontology" with a capital O means the discipline of metaphysics which consists of "study of being" more generally "study of existent." The appearance of this term in the field of computing dates back to the 1970s with the work of John McCarthy in the field of artificial intelligence [1]. During the 1990s, it was taken over in the field of knowledge representation. Today, the notion of ontology has a growing interest and has given birth to a whole discipline "ontological engineering." In fact, ontologies are now at the core of information retrieval or decision support systems in many fields such as health [2], legal [3], and learning [4].

Since then, the definitions of an ontology in computer science have multiplied. The most famous is Gruber's definition that defines an ontology as the "explicit specification of conceptualization" [5]. Then, Studer et al. have improved this definition by stating that an ontology is a "formal and explicit specification of a shared conceptualization" [6]. This definition puts forward four fundamental notions of ontology, namely:

- Formal specification: expressed in logical form, readable and understandable by machines, thus favoring unambiguous interpretations, which excludes the use of natural language.
- Explicit specification: explicitly defines the concepts and axioms that bind them.
- Conceptualization: abstract and simplified representation of the studied domain. According to Lee et al., a conceptualization is "an extraction of a domain vocabulary and is an abstract and simplified view of the domain that we wish to represent" [7].
- Shared: it describes consensual knowledge. In fact, ontology does not represent the viewpoint of an individual, but it serves as a general consensus, accepted and shared between the individuals of a whole community.

In addition, we are currently witnessing the third industrial revolution: the digital revolution. The changes brought by this revolution are rapidly and radically transforming the modern era. The field of data analysis does not escape the explosive rise of this revolution. This is due, on the one hand, to the technological evolution accomplished in the field of computers, on the other hand, to the progress made at the level of information processing models such as business intelligence [8], the artificial intelligence [9], and big data [10].

However, much of operational information system data in organizations remains unstructured and non-transactional. Indeed, according to Tseng & Chou [11], 20% of an organization's data is transactional, while the rest of the data is non-transactional and often saved in textual documents under an unstructured form. As a result, they remain unusable in an appropriate and effective way in the era of information capitalization. Thus, a phase of valorization of this data consists of extracting a knowledge base in the ontological form. The manual process of building such a base is laborious, tedious and time-consuming. In order to support this construction, several research projects have been carried out to automate the tasks involved. This automation is designated as 'Ontology Learning,' which refers to all the techniques and methods whose objective is the automatic or semi-automatic construction of ontologies via the extraction, the acquisition, and the generation from structured, semi-structured, or unstructured resources.

Tools to perform this extraction are numerous. Text2onto stands out as the one of the best tools known in this field. It has served as a subject of numerous studies of comparative evaluation of ontology learning tools in which it has excelled such as [12, 13]. The original version of Text2Onto is intended for the processing of the corpus written in English and is based on the Princeton WordNet (PWN) for English in order to perform some linguistic sub-processes such as the implementation of a variant of the algorithms intended for the discovery of Subclass of relationships. Hence, this tool is limited to the processing of textual resources written in English. To overcome this problem, we present in this paper our approach for the customization and adaptation of the English language version of the open-source Text2Onto tool to support the extraction of ontologies from French documents.

2.   ONTOLOGY LEARNING

Ontology notion is at the heart of the semantic web domain. We note a proliferation of research addressing the notion of ontology as illustrated in Figure 1 that gives an overview of statistics related to articles indexed in "Scopus.com" containing the term 'ontology.' These works cover one or more phases of the management of an ontology, such as the construction of an ontology, its storage, and its exploitation. The construction of an ontology can be classified into two families: manual construction and semi-automatic/automatic construction. The manual construction of an ontology is characterized by its reliability. Nevertheless, it is very tedious, very expensive and requires more time.

On the other hand, the second family, referred as 'Ontology Learning,' refers to all approaches whose objective is the semi-automatic construction of ontologies via the extraction, the generation and the acquisition. As part of Ontology Learning, several approaches have emerged to promote the accompaniment and facilitation of this painful construction. Among these approaches we distinguish that of ontology building from textual documents, dictionaries, knowledge bases, semi-structured data and relational databases. So, these approaches are variant primarily according to the data sources considered.

Several methodologies have emerged from the perspective of defining a roadmap for such a construction, among the best-known are Methontology [14], On-To-Knowledge [15], DILIGENT [16] and NeOn [17]. The construction of an ontology brings together a considerable set of tasks. The automation of this construction amounts to automating tasks or some of these tasks. The approaches for ontology construction from textual documents are grouped into several categories. Among others, we mention approaches based on linguistic techniques, on statistical techniques and those based on learning concept.

In this context, one of the most popular tools of ontology learning is Text2Onto [18]. This tool has served as several research works. In fact, on Google Scholar, the search of the term 'Text2onto' gives result including 988 pages displaying 10 elements, taking into consideration that we have excluded patents and

citations. This gives a result of 988*10 = 9880 (search date 05/12/2018) where the term 'Text2onto' appears. In addition, the Scopus indexing database lists 397 documents containing the term Text2Onto.

Moreover, our choice of the Text2Onto tool is justified by its performance confirmed by several research works, such as Toader Gherasim [12] work who made a comparative study of this tool with three other tools (OntoLearn, asium and sprat) and Jinsoo Park and al. [13] work who performed a comparison of the same tool with three other tools (DODDLE-OWL, OntoBuilder and OntoLT). Text2Onto tool provides to the designer, several algorithms dedicated to the different tasks of ontology generation [18]. The ontology designer chooses among these algorithms those to use for each task of this generation. The intervention of the designer allows the parameterization of the tool in addition to the refinement and validation of the intermediate result for the generation of the final ontology.

The operation of extracting an ontology from the text is based on analysis primitives. These primitives belong to the natural language processing and can be organized in a pyramidal model in the form of a 'Cake' as illustrated in Figure 2. The adaptation of Text2Onto tool for the French language goes, first of all, by the adaptation of these primitives.
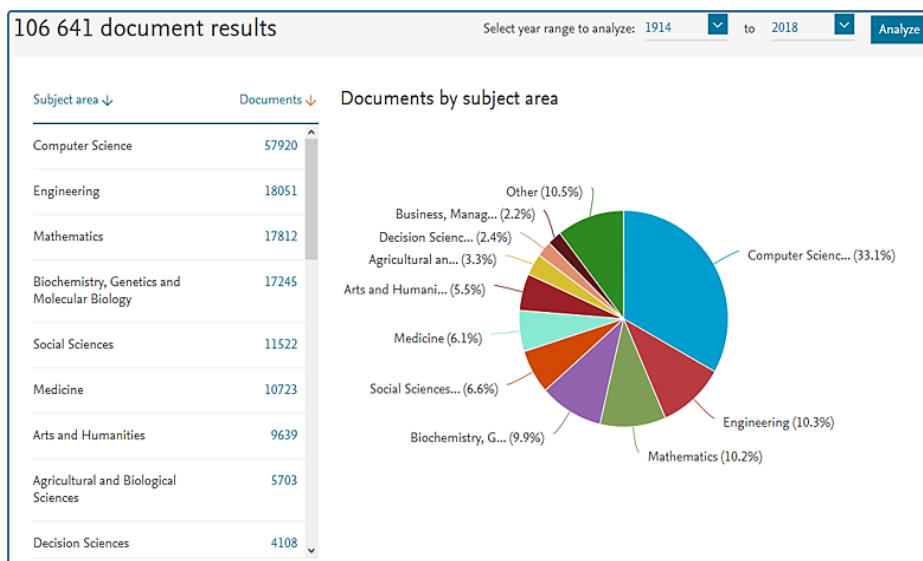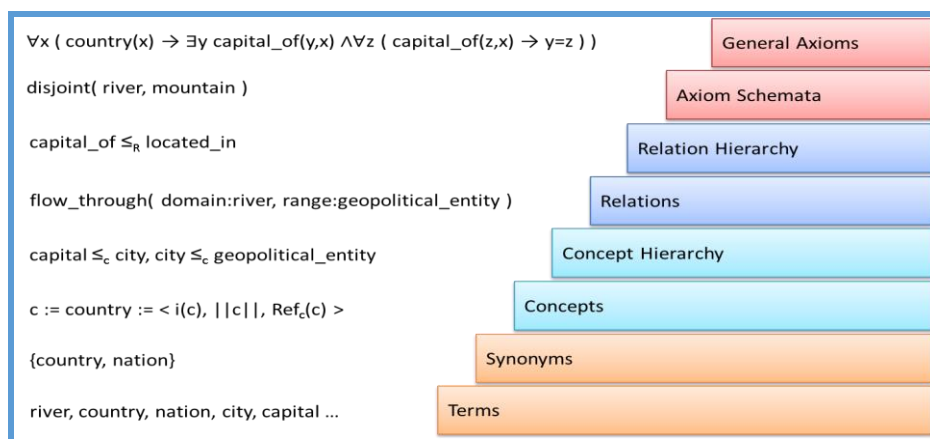

Figure 1. Ontology term proliferation

Figure 2. Ontology learning layer cake [19]

3. TEXT2ONTO

Text2Onto [18] is a framework for learning ontologies from textual data according to a transparent process. This tool has several algorithms grouped into seven sets such as "Concept," "Instance," "Similarity," "SubclassOf," "InstanceOf," "Relation," and "SubtopicOf." Among the algorithms of the group "SubclassOf"

is the algorithm WordNetConceptClassification that is based on the relationship of the structure of hypernym type in WordNet.

The designers of this tool implemented this algorithm in order to learn subclass-of relations by exploiting the hypernym relationship in WordNet. However, WordNet is an electronic lexical database for English. This constitutes a limitation for dealing with other languages with this tool, especially for learning subclass-of relations related to French.

In fact, P. Cimiano and V. Johanna have developed Text2Onto to overcome the other tools limitations such as TextToOnto [20], ASIUM [21], Mo'k Workbench [22], OntoLearn [23] or OnnearLT [24]. Among these limitations is the dependency of a specific or proprietary ontological model, the neglect of the interaction with the end user and the lack of consideration of changes in the data source. To overcome the first limitation, P. Cimiano and V. Johanna relied on a level of abstraction governed by the representation of an ontological structure by a meta-model based on modeling primitives. This gives their model the ability to translate the learned ontology into RDFS, OWL or F-Logic.

To overcome the second limitation, the authors adopted a probabilistic ontology model (POM) by presenting the results to the end user with probabilities of confidence. From the moment the results are ordered by certainty with the possibility of limiting them to those exceeding a confidence threshold, the interaction with the end user is made effective. To overcome the third limitation, the developers of this tool have developed it by exploiting the Data-Driven design pattern. That grants the tool the ability to handle changes in a corpus adequately. As a result, corpus changes are detected and POM changes are calculated based on these changes without making a recalculation of the POM for the whole corpus.

3.1. Linguistic preprocessing

To extract ontologies from a corpus, Text2Onto's authors have relied on the combination of the machine learning and natural language processing techniques. Indeed, Text2Onto is based on the general architecture for text engineering framework (GATE), which gives this tool a certain flexibility for adaptation and customization. Preprocessing begins with the preliminary stages of the Linguistic Preprocessing in Text2Onto, namely tokenization and sentence splitting. Then, the result of this step is used by a tagger POS by assigning each token its syntactic category. After, the phase of this pretreatment is completed by a morphological analyzer and a stemmer that perform lemmatization or stemming. So, basic preprocessing involves the following three steps:

− Tokenization and sentences splitting.
− Attribution of syntactic categories.
− Lemmatization or stemming.

The aforementioned steps constitute the basic pretreatment after which the annotated result thus obtained undergoes processing by a Java Annotation Patterns Engine (JAPE). This GATE module is used for surface analysis and identification of modeling primitives.

3.2. Concepts and instances

In the context of ontology learning, the extraction of terms that can represent concepts and relationships is based on two types of approaches: statistical approaches and linguistic approaches. Linguistic approaches are based on the analysis of the grammatical structure of the sentences that make up the text and on the identification of the syntactic (subject, direct object, etc.) and morpho-syntactic (noun, verb, adjective) roles of each word or group of words. While statistical approaches are based on calculating the relevance of a term in a corpus. In this context, Text2Onto developers have implemented algorithms to calculate:

− Relative term frequency (RTF).
− Term frequency inverted document frequency (TFIDF).
− Entropy.
− C-value/NC-value.

These values are normalized for each term and used as probability at the level of the POM.

3.3. SubclassOf relationships

In the original version of Text2onto, the developers of this tool implemented three algorithms to extract SubclassOf relationships. Indeed, they implemented the pattern concept classification, vertical relations concept classification and word net concept classification algorithms according to the approach proposed by P. Cimiano et al. [25]. These algorithms are used for extracting SubclassOf relationships from text written in English. For the support of French, we had to implement variants of these three algorithms adapted to the specifications of the free word net for French (WOLF) and the particularities of French.

- The implementation of the pattern concept classification algorithm is based on the techniques or methods of the natural language models, in particular on the Hearst-Patterns, illustrated in Figure 3, which is based on lexico-syntactic models based on hyponym and hypernym relationships.


```
(1) NP such as {NP, NP, ..., (and\or)} NP
(2) such NP as{NP,}* {(or\and)} NP
(3) NP {, NP}* {,} or other NP
(4) NP {, NP)* {,} and other NP
(5) NP {,} including {NP,}* {or\and} NP
(6) NP {,} especially {NP,}* {or\and} NP

NP stands for a noun phrase
```

Figure 3. Hearst patterns

- The implementation of the VerticalRelationsConceptClassification algorithm is based on the heuristic of "vertical relations" [26]. This algorithm is used for composite noun phrases.
- The implementation of the WordNetConceptClassification algorithm is based on the hypernym relationships from Princeton WordNet.

The standardized frequency of occurrences of a model represents the certainty of a relationship, in other words, the probability of a SubclassOf relationship.

3.4. Instance of relations

The creators of the Text2Onto tool have adopted an approach based on the notion of similarity. In the first place, they establish the computation of the contextual vectors for the instances as for the concepts. In the second place, they establish the computation of the similarity between the vectors of the instances and those of the concepts by using the divergence of Skewed [27]. Finally, they assign instances to the concepts with the greatest similarity as in [23].

4.   ADAPTING TEXT2ONTO FOR THE SUPPORT OF THE FRENCH LANGUAGE

In order to adapt the source code of the Text2onto tool, theoretical challenges and technical constraints must be taken into account. Indeed, each language is characterized by its specific lexical and syntactic morphologies. We begin the natural language processing of a corpus of documents in French language by pretreatment for elementary decomposition whose purpose is the extraction of the terms that make up the sentences.

This phase considers that a sentence consists of symbols whose union represents the text as a set. The operation of this splitting of the sentences of the text into symbols is referred by the tokenization. For this, we have exploited, customized and adapted the processing resources offered by GATE in this area. In fact, GATE has several types of resources, including processing resources, language resources, and visual resources. In order to reach this phase, we have used and parameterized the default tokeniser, the default gazetteer and the SentenceSplitter resources from the collection of reusable objects for language engineering plugin (CREOLE).

The adaptation of the CREOLE tools of treatment requires the customization of their parameters. Indeed, the tokenization operation refers to the fragmentation of the text into tokens using grammatical rules. These tokens represent the words, numbers and punctuation constituting the elementary building blocks with which the text is constructed. In the context of this paper, we have implemented and used a French Tokeniser.rules file listing the grammatical rules for the tokenization operation of text written in French. These rules are composed of two parts: the left part consists of regular expressions of correspondence and the right part gives annotations to assign. The gazetteer is used to annotate the proper names of countries, provinces, cities, organizations, days of the week, etc. based on files containing lists of these names. In this case, the file 'country.lst' lists the names of the countries, in our context, written in French. The gazetteer uses an indexing file 'lists.def' to access all the files in the lists.

In order to segment a text into sentences, the exploitation of the splitting operation based on the 'Sentence Splitter' resource offered by GATE is used. This resource consists of a suite of finite state transducers. This processing resource uses gazetteer lists and JAPE patterns defining the grammatical rules for sentence morphologies that we have established for sources written in French. GATE provides a syntax-based tagger (POSTagger: Part-Of-Speech Tagger) for identifying, independently of the language,

the syntactic category as well as the lemma of a given token. As part of GATE, tree tagging is done using specific shell scripts named treeTaggerBinary. To execute these scripts, we use a Java wrapper shell script.

We have developed algorithms for the processing of textual resources in French, including the WolfNet concept classification algorithm whose purpose is to extract relationships of SubclassOf type. Indeed, in order to learn subclass-of relations for text in French, we implemented in Text2Onto variants of algorithms exploiting the hypernym structure in WOLF. To be able to exploit this hypernym structure in WOLF, we have developed JWOLF [28] an API in Java based on the Java Architecture for XML Binding (JAXB) API for the manipulation of XML files.

5.   DISCUSSION

Initially, the adaptation of the Text2Onto tool for the processing of corpora of documents in French seemed an easy task. However, as we progress in this adventure, endless technical challenges arise continuously. We faced endless challenges that required painstaking efforts. It turns out that based on GATE, Text2Onto is closely related to this architecture. This strong dependency requires the learning and assimilation of concepts and notions on which GATE is built. The first lesson we learned is that sometimes it's better to build from the beginning than to try to adapt a pre-existing solution. In particular, if the solution no longer has support or at least contacts likely to shed light on the adaptation approach. The second lesson is that before attacking the personalization of a solution, we should master the theoretical notions and the technical tools on/with which it is established.

Armed with a great tenacity and determination, we have bravely managed the adaptation challenge of Text2Onto tool. At the end of this adventure, we managed to produce a version of Text2Onto able to generate an ontology from textual resources written in French. As it is shown in Figure 4, the results we have obtained are incredible and exceed our expected expectations. Indeed, by using this version for a simplistic document, we found that it allows the discovery of relationships whose existence is unsuspected and implausible.
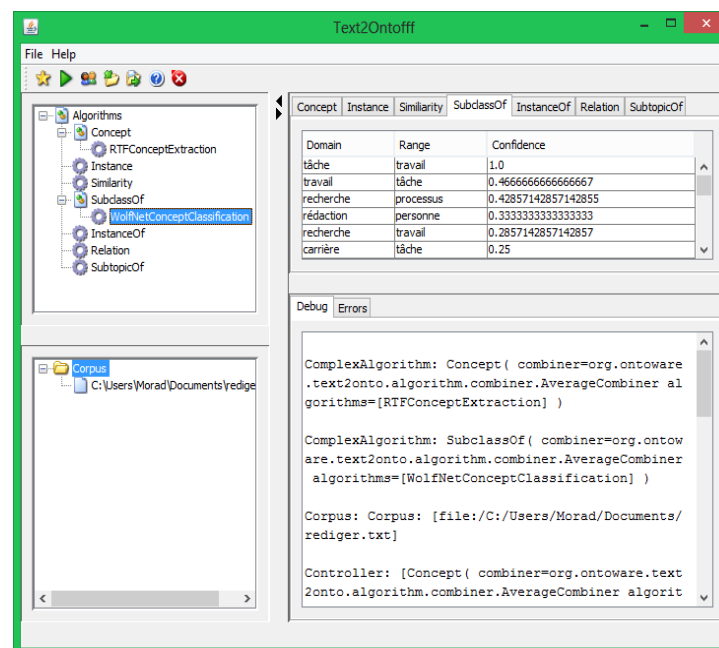

Figure 4. Text2Onto user interface–example of use for French

Figure 4 shows the results obtained for a simplistic example of a document written in French. Primarily, it illustrates the SubclassOf relationships obtained through the application of our WolfNet concept classification algorithm. For the development of this algorithm, we have based on the exploitation of the WOLF lexical database using the JWOLF API. Like its counterpart in the original version, we have invested the WOLF exploration features offered by JWOLF particularly in terms of structure investigations of the hypernym relationship between the different terms.

In Figure 5, we give the graphical representation of the instances of the SubclassOf relation extracted and illustrated in Figure 4. The most surprising is the ability of this version to discover instances of this relationship that we did not even imagine so that we relied on the features offered by the WOLF Browser [28] to ensure the veracity of the results. At this level, we should expect the unexpected.
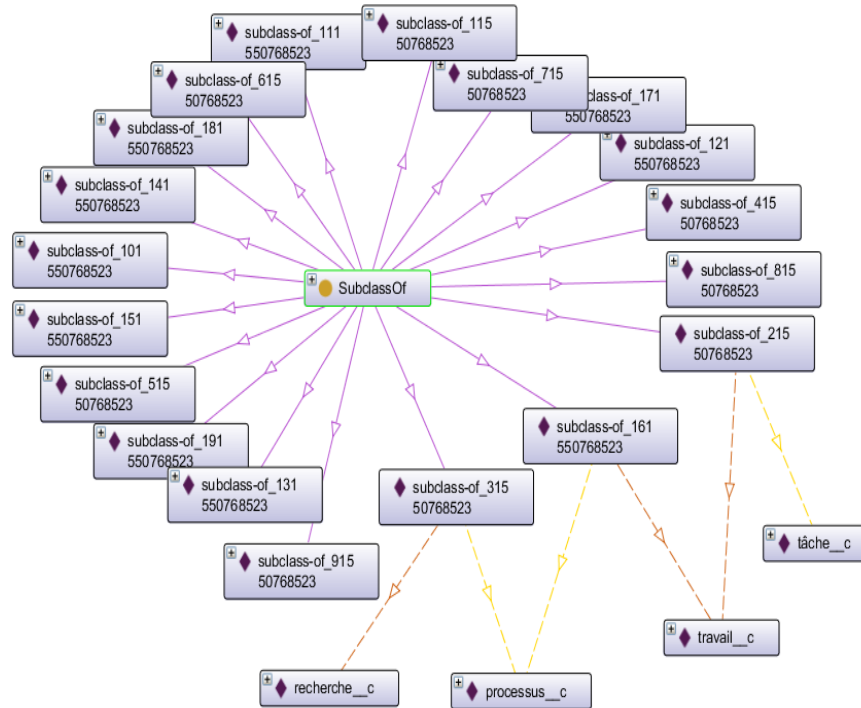

Figure 5. Graphical view of SubclassOf relationship instances

6. CONCLUSION

In this paper, we presented our approach for adapting the Text2Onto tool for the extraction of ontologies from corpora of textual documents written in French. Despite the obstacles encountered due to lack of documentation and support, we managed to produce an enhanced version of Text2Onto able to generate an ontology from a corpus in French. Admittedly, with this version we are able to extract such an ontology. However, the adaptation of this tool is extremely overwhelming. Indeed, this version allows us to extract an ontology from a corpus of documents in French, nevertheless the performances do not satisfy our perfectionist expectations. However, this adventure has opened the way to set up a more sophisticated version.

In future work, more efforts are needed, on the one hand, to further improve the algorithms that we have adapted, on the other hand, to optimize the execution time and utilization of resources in terms of CPU consumption and memory occupation. Not to mention, evaluating this version on a large scale. The result of this contribution is the broadening of our horizons in the field of ontology learning. One of the auxiliary results is the investigation of the bases on which Text2Onto is built and the technical and methodological aspects of its development. This is an enlightened and informed way opening for the adaptation of this tool to support the other languages.

REFERENCES

[1] V. Psyche, J. Bourdeau, and O. Mendes, "Apport de l'ingénierie ontologique aux environnements de formation à distance," *Sciences et Technologies de l'Information et de la Communication pour l'Éducation et la Formation*, vol. 10, no. 1, pp. 89–126, 2003.

[2] X. Aimé, J. CharletJ, F. Furst, P. Kuntz, F. Trichet, and F. Dhombres, "Rare diseases knowledge management: the contribution of proximity measurements in OntoOrpha and OMIM," *Studies in Health Technology and Informatics*, pp. 88–92, 2012,

[3] N. Casellas, *Legal ontology engineering: Methodologies, modelling trends, and the ontology of professional judicial knowledge*, vol. 3. Springer Science & Business Media, 2011.

[4]　M. HAJJI, M. QBADOU and K. MANSOURI, "Towards an Automatic Authoring and Optimization System of Adaptive Course Materials," *2019 5th International Conference on Optimization and Applications (ICOA)*, Kenitra, Morocco, 2019.

[5]　T. R. Gruber, "A translation approach to portable ontology specifications," *Knowledge Acquisition*, vol. 5, no. 2, pp. 199–220, Jun. 1993.

[6]　R. Studer, V. R. Benjamins, and D. Fensel, "Knowledge engineering: Principles and methods," *Data & Knowledge Engineering*, vol. 25, no. 1, pp. 161–197, Mar. 1998.

[7]　J. Lee, H. Chae, C.-H. Kim, and K. Kim, "Design of product ontology architecture for collaborative enterprises," *Expert Systems with Applications*, vol. 36, no. 2, Part 1, pp. 2300–2309, Mar. 2009.

[8]　S. Negash and P. Gray, "Business Intelligence," in *Handbook on Decision Support Systems 2: Variations*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 175–193.

[9]　D. McDermott, "Artificial Intelligence Meets Natural Stupidity," *SIGART Bull.*, no. 57, pp. 4–9, Apr. 1976.

[10]　De Mauro Andrea, "A formal definition of Big Data based on its essential features," *Library Review*, vol. 65, no. 3, pp. 122–135, Jan. 2016.

[11]　F. S. C. Tseng and A. Y. H. Chou, "The concept of document warehousing for multi-dimensional modeling of textual-based business intelligence," *Decision Support Systems*, vol. 42, no. 2, pp. 727–744, Nov. 2006.

[12]　T. Gherasim, "Détection de problèmes de qualité dans les ontologies construites automatiquement à partir de textes," Intelligence artificielle [cs.AI], Université de Nantes, 2013.

[13]　J. Park, W. Cho, and S. Rho, "Evaluating ontology extraction tools using a comprehensive evaluation framework," *Data & Knowledge Engineering*, vol. 69, no. 10, pp. 1043–1061, 2010.

[14]　M. Fernández-López, A. Gómez-Pérez, and N. Juristo, "METHONTOLOGY: From Ontological Art Towards Ontological Engineering," *AAAI-97 Spring Symposium Series*, pp. 33–40, 1997.

[15]　S. Staab, R. Studer, H.-P. Schnurr, and Y. Sure, "Knowledge processes and ontologies," *IEEE Intelligent Systems*, vol. 16, no. 1, pp. 26–34, Feb. 2001.

[16]　H. S. Pinto, S. Staab, and C. Tempich, "DILIGENT: Towards a fine-grained methodology for DIstributed, Loosely-controlled and evolvInG Engineering of oNTologies," i*n Proceedings of the 16th European Conference on Artificial Intelligence*, pp. 393–397, 2004.

[17]　M. C. Suárez-Figueroa, "NeOn Methodology for Building Ontology Networks: Specification, Scheduling and Reuse," Jun-2010.

[18]　P. Cimiano and J. Völker, "Text2Onto: A Framework for Ontology Learning and Data-Driven Change Discovery," in *Lecture Notes in Computer Science*, vol. 3513, pp. 227–238, 2005.

[19]　P. Cimiano, "Ontology Learning and Population from Text: Algorithms, Evaluation and Applications," *Berlin, Heidelberg: Springer-Verlag*, 2006.

[20]　A. Maedche and S. Staab, "Discovering Conceptual Relations from Text," in *Proceedings of the 14th European Conference on Artificial Intelligence*, Amsterdam, The Netherlands, The Netherlands, pp. 321–325, 2000.

[21]　D. Faure and C. Nédellec, "A Corpus-based Conceptual Clustering Method for Verb Frames and Ontology Acquisition," in *LREC workshop on*, pp. 5–12, 1998.

[22]　G. Bisson, C. Nédellec, and D. Cañamero, "Designing Clustering Methods for Ontology Building - The Mo'K Workbench.," in *ECAI'2000 Workshop on Ontology Learning, Proceedings of the First Workshop on Ontology Learning OL'2000, Berlin, Germany, August 25, 2000. Held in conjunction with the 14th European Conference on Artificial Intelligence ECAI'2000, Berlin, Germany*, 2000.

[23]　E. Alfonseca and S. Manandhar, "Extending a Lexical Ontology by a Combination of Distributional Semantics Signatures," in *Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web*, Berlin, Heidelberg, pp. 1–7, 2002.

[24]　J. Charlet, B. Bachimont, and R. Troncy, "Ontologies pour le web sémantique," *Le Web sémantique - Revue I3 : Information - Interaction - Intelligence - Numéro hors série*, pp. 69–100, 2004.

[25]　P. Cimiano, A. Pivk, L. Schmidt-Thieme, and S. Staab, *"*Learning Taxonomic Relations from Heterogeneous Sources of Evidence," *IOS Press Style Sample*, pp. 1-15, 2019.

[26]　P. Velardi, P. Fabriani, and M. Missikoff, "Using text processing techniques to automatically enrich a domain ontology," in *Proceedings of the international conference on Formal Ontology in Information Systems - Volume 2001*, Ogunquit, Maine, USA, pp. 270–284, 2001.

[27]　L. Lee, "Measures of Distributional Similarity," in *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, Stroudsburg, PA, USA, pp. 25–32, 1999.

[28]　M. Hajji, M. Qbadou, and K. Mansouri, "JWOLF: Java Free French Wordnet Library," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 2, pp. 339-345, 2019.