International Journal of Engineering Trends and Technology
ISSN: 2231–5381 / https://doi.org/10.14445/22315381/IJETT-V73I1P106

Volume 73 Issue 1, 77-85, January 2025
© 2025 Seventh Sense Research Group®

Original Article

Domain Ontology Extraction from a Glossary: Case of the Phosphate Industry

Oussama Chabih[1], Sara Sbai[2], Mohammed Reda Chbihi Louhdi[3], Hicham Behja[4]

[1, 2,4]LRI – Laboratory, ENSEM, Hassan II University, Casablanca, Morocco.
[3]LIS – Laboratory, Faculty of Sciences Ain Chock, Hassan II University, Casablanca, Morocco.

[1]Corresponding Author : chabih.oussama@gmail.com

Received: 04 September 2024 Revised: 06 November 2024 Accepted: 27 December 2024 Published: 31 January 2025

Abstract - *Ontologies play a crucial role in structuring domain-specific knowledge, enabling more efficient search, discovery, and data interoperability across different systems. In this context, an innovative approach to transform domain-specific glossaries into ontologies, with a focus on the phosphate industry, is proposed in this paper. Unlike traditional methods that depend solely on transformation rules, the proposed approach combines empirical and algorithmic techniques to detect relationships between glossary terms, resulting in a more accurate and comprehensive ontology. The proposed method was applied to the OCP Group's internal glossary, successfully generating an OWL2 ontology that significantly improves search and discovery within the organization's knowledge management systems. The experiment's results demonstrate that the proposed method outperforms existing techniques in terms of accuracy and relevance, providing a robust framework for knowledge representation in specialized industrial contexts.*

Keywords - *OWL2 Ontology, Ontology extraction, Transformation rules, Jaccard similarity, Glossary vocabulary.*

1. Introduction

The transformation of data into ontologies is a critical task in the field of information management, particularly for enhancing data search, exploration, and interoperability. Ontologies offer a formal structure for representing knowledge, enabling more effective data usage in various applications. This paper proposes an innovative approach specifically tailored to the needs of the OCP Group, a phosphate industry leader. The approach involves the extraction of an OWL2 ontology from a specialized glossary, utilizing a novel combination of domain expertise and Natural Language Processing (NLP) techniques. While considerable research has been devoted to extracting ontologies from structured data sources like relational databases and semi-structured data such as JSON and XML, the number of studies focusing on ontology extraction from glossaries remains quite modest. The existing methods that address glossary-based ontology extraction typically rely on straightforward transformation rules, often without robust mechanisms for detecting relationships between concepts, or they offer only very limited and direct relationship detection. The motivation for generating this ontology stems from the specific needs of the OCP Group to manage and utilize its vast repository of specialized knowledge accumulated over decades. This knowledge, encapsulated in a domain-specific glossary, is critical for ensuring the efficiency and relevance of internal information retrieval systems. The aim is to enhance the group's ability to search and discover relevant information by transforming this glossary into a comprehensive OWL2 ontology, thereby improving decision-making processes and fostering innovation within the organization. To achieve this, a dual-method approach that combines empirical techniques for reusing keywords with algorithmic methods, specifically the Jaccard similarity, to detect and structure relationships between glossary terms is introduced. This approach ensures that the resulting ontology is both accurate and tailored to the specialized language and concepts of the phosphate industry. The rest of this paper is organized as follows: Section II reviews related works in the area of ontology extraction, focusing on methods applied to glossaries. Section III details the proposed method for extracting an OWL2 ontology from the OCP Group's glossary, including the specific transformation rules and relationship detection techniques employed. Section IV presents the experimental setup and results, demonstrating the effectiveness of the proposed approach compared to existing methods. Finally, Section V concludes the paper by discussing the contributions, potential applications, and future work in this area.

2. Related Works

Much work has been undertaken to transform different types of data sources into ontologies, with the aim of leveraging the benefits they offer. Ontologies, in fact, simplify the search and discovery of information by making it possible to formalize and represent knowledge in a precise and structured manner. They define concepts as well as the relationships and constraints that govern them.

Additionally, ontologies promote data interoperability, making it easier to exchange data between different systems. They are widely used in automated reasoning to deduce new information from existing knowledge in the ontology. By structuring data and establishing semantic relationships between concepts, ontologies significantly improve the precision and relevance of information searching on the web. In short, their use helps to make data more accessible, understandable and usable in various online contexts and applications. This section reviews existing literature and research efforts related to developing, applying, and evaluating ontologies across various domains. The objective is to identify gaps in current knowledge and suggest areas for future exploration.

Various approaches have been developed for constructing ontologies utilizing different data sources. For relational databases, Benamar et al. [1] proposed a Model-Driven Engineering (MDE) approach to convert SQL data into an OWL ontology, while Lakzaei and Shamsfard [2] used mapping rules to transform a normalized relational database into an ontology. Additionally, Ben Mahria et al. [3] introduced a method to extract ontologies from SQL files, generating both concepts (T-Box) and instances (A-Box) for the final OWL ontology. OntoBase [4] automates ontology creation from databases using the Protégé API, allowing new classes to be derived from table columns and database schemas to be converted into database concepts. Despite advancements, challenges remain in accurately mapping foreign keys and translating properties, as noted in the studies cited.

Efforts have also been made to extract ontologies from NoSQL databases. Curé et al. [5] developed a framework for integrating data from MongoDB and Cassandra into an ontology through three steps: creating ontologies for each data source, aligning these ontologies into a global one, and processing SPARQL queries into Java code via an intermediate language, BQL. Similarly, Kiran et al. [6] proposed a semantic integration system for HBase, comprising schema generation, extraction and conversion of column details into ontology entities, alignment of these ontologies, and using the resulting global ontology as a T-Box for OWL reasoners querying a SPARQL endpoint. Abbes et al. [7] suggested an approach for building ontologies from MongoDB by defining transformation rules for ontology skeleton creation, identifying properties and data types, detecting individuals, and deducing axioms and constraints. In [8], Jabbari et al. employed Formal Concept Analysis (FCA) [9] and transformation rules to generate an ontology from a document-oriented NoSQL database. Ontology construction methods from semi-structured data are diverse and still evolving. Yao et al. [10] proposed a method for converting web data into semantic web descriptions using key-value pairs in JSON objects. This enables the creation of semantic models for data instances through four steps: JSON object analysis, semantic mapping, semantic enrichment, and ontology alignment. Users can then validate, modify, and apply the constructed ontologies. In [11], a Protégé plugin, OWLET, was introduced to aid experts during the refinement phase of ontology construction, helping transform real-world objects (like images) into instances for integration into existing ontology models. Baek et al. [12] outline a method for creating ontology knowledge bases from semi-structured datasets (e.g., spreadsheets, JSON, XML) by extracting target columns, utilizing a Transform Table Generator (TTG) and Cell Value Importer (CVI) to import values, and applying property expressions (PropertyExp) to map columns to properties. Seidel et al. [13] proposed a tool for extracting knowledge from heterogeneous semi-structured data sources, involving source file preparation for annotation, dictionary analysis of values, and resolution of multiple references to map predicates. Few studies have been identified for the construction of ontologies from glossaries. In [14], the paper presents a method for extracting structured knowledge for cultural heritage applications. The approach involves pre-processing, where terms and definitions from glossaries are processed using part-of-speech tagging and Named Entity Recognition, then annotation, where glossary definitions are segmented and annotated with semantic properties using a model like CIDOC-CRM and finally, Ontology formalization, where the annotated segments are transformed into formal semantic structures. In another work [15], the authors propose the creation of a disaster management domain from five different emergency management glossaries and vocabularies.

In [16], the authors propose a methodology to create a software testing ontology from a glossary, using the ONTO6 approach to identify the essential aspects of the domain. This method includes several steps, including extracting key concepts, creating an aspect graph representing the ontology, and iterating to refine the results with expert input. It aims to provide a clear structure for understanding knowledge in software testing, thus facilitating access to relevant information and creating more effective testing tools. In another article [17], the authors presented a method for ontology development in seven steps, ranging from domain determination to instance creation. This method includes the creation of an XML dictionary, transformation into a taxonomy with hierarchical relationships, and then into a thesaurus with equivalent and associative relationships. Finally, it results in a semantically enriched ontology with properties and restriction rules based on a linguistic analysis of the terms. Another method [18] of ontology construction offers a systematic approach to clarify the meaning of Web resources, thus facilitating their processing by machines. This approach is based on a representation structure called "extended glossary language" (ELL), allowing precise modelling of concepts. By following key steps such as term clustering, relationship identification, and synonym disambiguation, this method aims to simplify and streamline the process of building ontologies, thereby making web resources more accessible and understandable for software agents. In [19], Loris et al. described a nine-phase methodology for developing an ontology from an initial glossary. This method includes grouping terms into clusters, adding new terms found in definitions, identifying relationships between terms, disambiguating synonyms, grouping classes, conceptual modelling, schema representation, the representation of the ontology, and finally, the annotation of classes and individuals with information derived from the glossary. Each phase contributes to progressively building the structure and semantics of the ontology.

In [20], a method was presented to describe the process of developing an ontology from a semi-structured glossary, using natural language processing techniques to extract essential aspects of the domain. This method offers a semi-automatic transformation of a glossary into a navigable concept map, allowing learners to explore concepts and their definitions dynamically. It also highlights the use of glossaries in ontology construction, offering a nine-phase methodology for this process, highlighting the elucidation of significant aspects of the domain and the creation of aspect graphs to facilitate understanding of the domain studied. Finally, in [21], the authors introduce an ontology-driven approach to integrate heterogeneous databases, employing natural language processing and semantic modelling techniques. The constructed knowledge graph supports applications like decision-making and operational monitoring in the petroleum industry. The domain ontology creation uses petroleum exploration glossaries. These studies collectively showcase the diverse methodologies and innovations in the field of semantic search engines and information retrieval, providing a rich background for the current research.

3. Proposed Method

The primary challenge in extracting ontologies from domain-specific glossaries lies in accurately capturing and representing the intricate relationships between specialized terms. Traditional methods often fall short in industries like the phosphate sector, where the glossary reflects decades of accumulated knowledge and industry-specific terminology. These methods typically focus on simple transformation rules that convert glossary terms into ontological components, but they lack robust mechanisms to detect and structure the relationships between these terms.

Existing ontology extraction techniques from glossaries are limited in several ways:

- Limited Scope: Most existing methods offer only basic transformation rules without delving into the relationships between concepts. This results in overly simplistic ontologies that fail to represent the complex interdependencies inherent in specialized domains.
- Direct and Simplistic Relationship Detection: The few methods that attempt to identify relationships often rely on direct or superficial techniques, which can miss more nuanced or implicit connections between terms.
- Lack of Adaptation to Domain-Specific Needs: These methods are not tailored to specific industries' unique linguistic and conceptual characteristics, leading to ontologies that may not fully capture the domain's complexity.

Within the OCP Group's research and innovation department, the legacy of forty years of activity results in an invaluable wealth of information, mainly summarized in numerous reports. Faced with this valuable but sometimes complex reservoir, the objective of optimizing access and use of this data has led to the design of an innovative internal search engine architecture. To achieve this objective, the proposed approach is initiated by analysing the various documents circulating within the company. A company glossary was created from these documents and has improved over the years. This glossary is a mixture of definitions, abbreviations, and chemical formulas, either existing and defined from the department's point of view or innovative, resulting from years of research.

This makes the task of finding a definition of these types of terms elsewhere than in the internal glossary difficult. The majority of keywords within the glossary pertain to chemistry and physics. Numerous terms have been either invented by the company or used in ways that differ from common scientific language. This complexity poses a challenge when employing a general phosphate ontology, potentially diverting research away from the company's specific needs. Consequently, it was decided to generate an ontology directly from the company’s glossary, enhancing it with content from articles found on the company's website and leveraging the expertise of the company's professionals. This section delves deeper into the proposed method for emerging an ontology designed from the company’s glossary and the rich range of accumulated reports.

3.1. Methodology for Ontology Construction

According to Tom Gruber [22], an ontology is an "explicit specification of a conceptualization of a domain of interest," while Swartout and colleagues [23] define it as "a hierarchically structured set of terms to describe a domain, which can serve as a basis for a knowledge base." Ontology refers to the science of describing the various types of entities in the world and their relationships. In the web context, it defines the terms used to describe an area of expertise, represented by structured diagrams that are readable by computers. By facilitating interoperability between systems, an ontology can be compared to a database but with a vast network of relationships between concepts. The benefits of using ontologies include improved web searchability, enhanced knowledge sharing, the ability to reuse knowledge across different domains, and easier adaptation to changes within a domain. To ensure the development of a robust ontology, specific guidelines rooted in the glossary analyses were established. These guidelines focus on accurately identifying key concepts, creating corresponding OWL classes, and defining relevant properties, all tailored to the specific language and terminologies used within the OCP Group:

- Identification of Concepts: The glossary was scrutinized to identify key concepts, including different acids (e.g., Sulfuric Acid, Hydrofluoric Acid), along with their associated relationships and properties.
- Creation of Classes: Each identified concept was declared as a class in OWL. For instance, "Acide" (English: "Acid") was represented by a class `ont:Acide`.
- Definition of Properties: Relationships between concepts were represented by properties. For example, the relationship "contient" (English: 'contains') was implemented as the property `ont:contient`.
- Creation of Annotation Properties: Unique properties were added to each class to represent distinct information. For instance, `Formule_chimique` for the chemical formula, 'definition' for keyword definitions, and 'abréviation' for keyword abbreviations.
- Creation of Named Individuals: Each specific element in the glossary (e.g., "Sulfuric Acid in French Acide sulfurique") was represented as a named individual, linked to its corresponding class (`rdf:type ont:AcideSulfurique`).
- Addition of Properties to Individuals: Specific properties were appended to each individual to provide additional details, such as the chemical formula of a particular acid.
- Relationship Management: Subclasses and equivalent classes were managed using `rdfs:subClassOf` and `owl:equivalentClass`, respectively. For instance, different acids were considered subclasses of the 'Acid' class, while synonyms were treated as equivalent classes.
- Finalization of the Ontology: The entire ontology underwent a comprehensive review to ensure the consistency of relationships, properties, and classes. Adjustments were made as needed.
- Verification with an OWL Tool: The Protégé tool was employed to create, visualize, and validate the ontology, ensuring its compliance with OWL standards.

Given these challenges, the approach is designed to address the limitations of traditional methods by combining empirical and algorithmic techniques. Initially, a set of transformation rules specifically tailored to the language and terminology used within the OCP Group are applied. These rules guide the conversion of glossary terms into foundational ontological components, such as classes and properties, ensuring that the resulting ontology reflects the domain's complexity and nuances.

3.2. Relationship Between Concepts

The first phase of the proposed methodology involves an empirical approach that builds on these transformation rules by leveraging domain-specific knowledge to hypothesize potential relationships between glossary terms. This approach consists of the following steps: the Empirical method, which leverages domain-specific knowledge through keyword reuse and definition analysis, and the Algorithmic method, which uses Jaccard similarity to quantitatively assess and validate relationships. This dual approach enables the capture of both explicit and implicit relationships, resulting in a more comprehensive and nuanced ontology.

3.2.1. Empirical Method

This phase leverages domain-specific knowledge to hypothesize potential relationships between glossary terms. It consists of the following steps:

- Keyword Reuse: searching for instances where one glossary term is used within the definition of another. For example, if the term "phosphoric acid" appears within the definition of "acid," this suggests a potential relationship between the two terms.
- Chemical Formula Analysis: Given the chemical nature of many terms in the glossary, chemical formulas were analyzed to detect relationships based on chemical composition. For instance, the relationship between sulfuric acid ($H2SO4$) and its components, such as sulfur trioxide ($SO3$) and water ($H2O$), can be identified through their chemical reaction.
- Definition Analysis: the definitions of terms are examined to identify cases where one term is used to define another, indicating a hierarchical or compositional relationship; for example, the keyword "Ammonium Sulfate Phosphate" is used in the definition of the keyword "Fertilizer".

These Empirical methods provide initial insights into potential relationships, which are validated and refined through more rigorous algorithmic analysis.

3.2.2. Algorithmic Method

To validate, refine, and detect more relationships identified by the empirical approach, an algorithmic method based on similarity with Jaccard is used. The selection of the Jaccard similarity over other metrics, such as cosine similarity or Levenshtein distance, is due to its suitability for the data and the goals of this research. Unlike cosine similarity, which measures the cosine of the angle between two vectors and is sensitive to the length and frequency of terms, Jaccard similarity focuses solely on the presence or absence of terms within sets.

This makes it particularly effective in this context, where the primary concern is the overlap of concepts between definitions rather than the frequency or direction of terms.

Additionally, while Levenshtein distance calculates the minimum number of single-character edits required to transform one word into another, it is less effective for comparing entire sets of words, especially in specialized domains with complex terminologies. Jaccard similarity, by contrast, provides a more straightforward and computationally efficient way to assess the similarity between the sets of words that constitute the definitions of different glossary terms, making it the most appropriate choice for detecting relationships in the ontology extraction process. The Jaccard method is a similarity measure used to compare the similarity between two sets, whether sets of words, characters, or other elements. This metric is widely used in the field of natural language processing and text data analysis. Jaccard's formula is defined as the ratio between the number of elements common to two sets and the total number of unique elements present in these sets. Mathematically, it is expressed as follows:

$$J(A,B) = \frac{|A \cap B|}{|A \cup \ B|}$$

Where:

- A and B are the two sets being compared.
- |A∩B| denotes the number of words common to both definitions A and B.
- |A∪ B| denotes the total number of unique words present in definitions A and B.

To categorize related suspicious classes based on the similarity results, the following ranges were defined:

- 0: No similarity.
- 0.01 to 0.3: Low similarity.
- 0.3 to 0.5: Moderate similarity.
- 0.5 to 0.7: Moderate to good similarity.
- 0.7 to 1: High to perfect similarity.

The last similarity range (0.7 to 1) is considered a probable relationship, with experts tasked to define the nature of the relationship. These chosen ranges adhere to a common convention in natural language processing, facilitating the qualitative analysis of semantic relationships within the ontology.

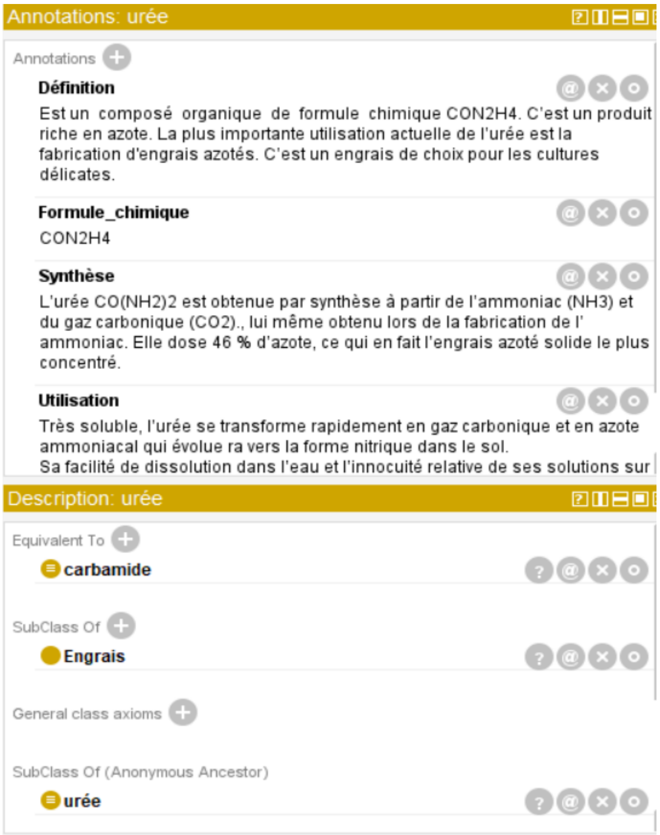
The steps involved in this phase are as follows:

- Text Preprocessing: Tokenization, stemming, and stop word removal are performed on the glossary text before applying algorithmic analysis. Additionally, keywords are unified by replacing abbreviations, synonyms, and chemical formulas with a standardized term.
- Jaccard Similarity Calculation: The Jaccard similarity is calculated between the definitions of different glossary terms. This metric is particularly suited to the data as it measures the ratio between the number of common elements (words) and the total number of unique elements in two sets. This approach allows the detection of relationships that are not immediately apparent through direct keyword matching.
- Classification of Relationships: Based on the Jaccard similarity scores, relationships are classified into categories ranging from "no similarity" to "high similarity." Relationships with a high similarity score are considered strong candidates for inclusion in the ontology, while those with lower scores may require further expert validation.

After applying empirical and algorithmic methods to the glossary, the results are presented to a domain expert within the company to validate the extracted potential relationships and define the nature of these relationships, thus enriching the obtained ontology with these validated relationships.

4. Results and Discussion

To evaluate the effectiveness of the proposed method, a Python program utilizing the RDFLib library was developed. This program enables the application of the proposed rules and methods to the case study—the internal glossary of the OCP Group. Additionally, the scikit-learn library was used for natural language processing tasks. First, the rules to define the concepts as well as their properties were applied. Next, natural language processing was performed to apply the proposed algorithms.



Annotations: urée

Annotations

Définition

Est un composé organique de formule chimique CON2H4. C'est un produit riche en azote. La plus importante utilisation actuelle de l'urée est la fabrication d'engrais azotés. C'est un engrais de choix pour les cultures délicates.

Formule_chimique

CON2H4

Synthèse

L'urée CO(NH2)2 est obtenue par synthèse à partir de l'ammoniac (NH3) et du gaz carbonique (CO2)., lui même obtenu lors de la fabrication de l' ammoniac. Elle dose 46 % d'azote, ce qui en fait l'engrais azoté solide le plus concentré.

Utilisation

Très soluble, l'urée se transforme rapidement en gaz carbonique et en azote ammoniacal qui évolue ra vers la forme nitrique dans le sol.
Sa facilité de dissolution dans l'eau et l'innocuité relative de ses solutions sur

Description: urée

Equivalent To

carbamide

SubClass Of

Engrais

General class axioms

SubClass Of (Anonymous Ancestor)

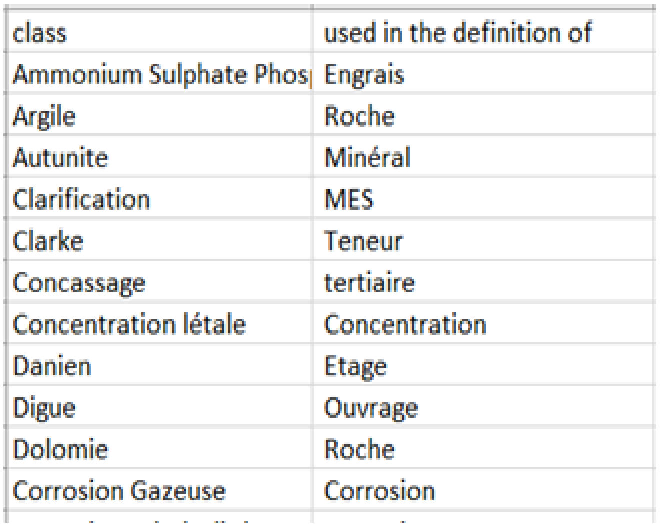
urée

Fig. 1 Example of a generated class generated

Figures 1 and 2 present an example of a class generated using the proposed approach. Regrettably, for confidentiality reasons, only limited data is presented in this work, primarily those found in standard glossaries. Figure 3 displays some of the keywords used to define other keywords. Overall, the use of 13.7% of the keywords in the definition of others was detected.

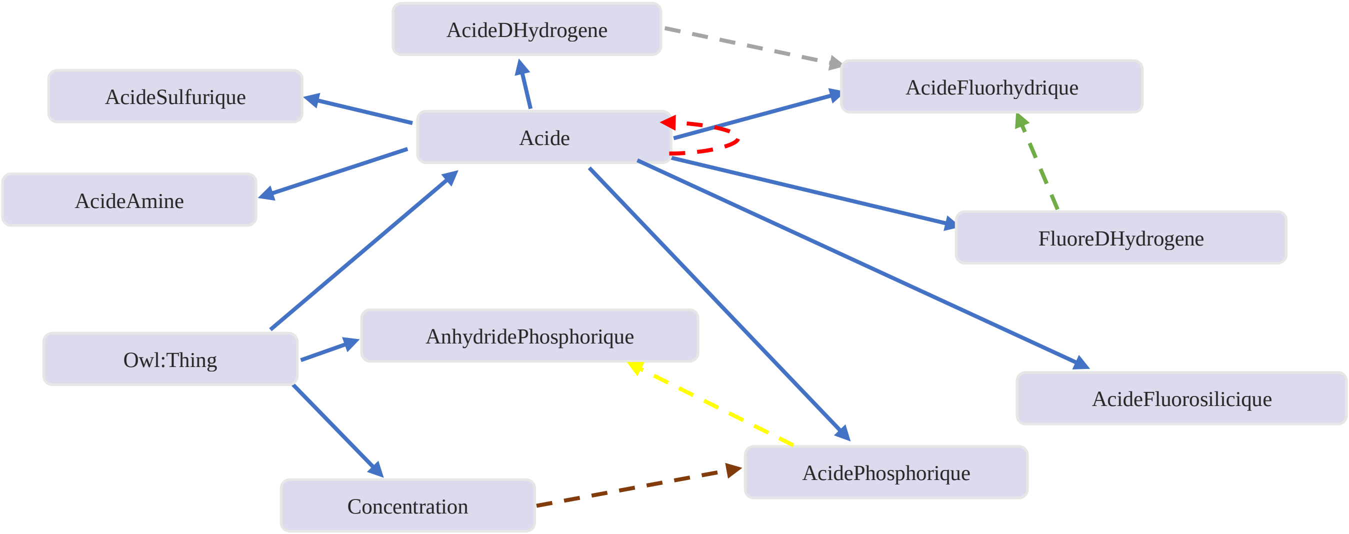
```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix ont: <http://www.Lexique.com#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
<http://www.Lexique.com#urée>
rdf:type ont:Class ;
rdfs:subClassOf <http://www.Lexique.com#Engrais> ;
ont:Définition "Est un composé organique de formule chimique CON2H4. C'est un produit riche en azote. La plus importante utilisation actuelle de l'urée est la fabrication d'engrais azotés. C'est un engrais de choix pour les cultures délicates."^^xsd:string ;
ont:Formule_chimique "CON2H4"^^xsd:string ;
ont:Synthèse "L'urée CO(NH2)2 est obtenue par synthèse à partir de l'ammoniac (NH3) et du gaz carbonique (CO2), lui-même obtenu lors de la fabrication de l'ammoniac. Elle dose 46 % d'azote, ce qui en fait l'engrais azoté solide le plus concentré."^^xsd:string ;
ont:Utilisation "Très soluble, l'urée se transforme rapidement en gaz carbonique et en azote ammoniacal qui évolue vers la forme nitrique dans le sol. Sa facilité de dissolution dans l'eau et l'innocuité relative de ses solutions sur le feuillage permettent de l'utiliser pour les pulvérisations foliaires (soit seule, soit en mélange avec des traitements antiparasitaires), et pour l'irrigation fertilisante."^^xsd:string .
```

Fig. 2 Example of a generated Class, Written in Turtle



class	used in the definition of
Ammonium Sulphate Phos	Engrais
Argile	Roche
Autunite	Minéral
Clarification	MES
Clarke	Teneur
Concassage	tertiaire
Concentration létale	Concentration
Danien	Etage
Digue	Ouvrage
Dolomie	Roche
Corrosion Gazeuse	Corrosion

Fig. 3 Some examples of classes used in defining others



– – –▶	**estUnComposantDe** (Domain > Range)
– – –▶	**estUnComposeDe** (Domain > Range)
– – –▶	**estUneConcentrationDe** (Domain > Range)
– – –▶	**estUneFormeDe2** (Domain > Range)
– – –▶	**estUneFormeDe** (Domain > Range)
——▶	**has subclass**

Fig. 4 Example of classes and relations generated by the proposed approach

Figure 4 shows the obtained ontology, highlighting the relationships between different classes of acids and their properties. The proposed approach allowed the detection and modelling of hierarchical and transitive relationships between classes. In this example, the inheritance relationship is defined as a hierarchical relationship between the different acid classes. For example, "Acide Sulfurique" in English "sulfuric acid," "Acide Phosphorique" in English "phosphoric acid," and "Acide d'Hydrogène" in English "hydrogen acid" are all subclasses of the "Acide" in English "Acid” class.

This hierarchical structure represents the parent-child links between the different types of acids. The relation "est Un Composant De" in English "is a compound of" is also a transitive relation that links classes together. For example, "Acide Phosphorique" (phosphoric acid) is a compound in the "Anhybride Phosphorique" (phosphoric anhydride) class. This relationship models how different acids are composed of more fundamental sub-elements.

Additionally, the "Concentration" class was introduced to represent the different concentrations of phosphoric acid. The relationship "est Une Concentration De" in English "is a concentration of" links this class to phosphoric acid, establishing a direct link between the concentration and the main chemical component. Although the notion of concentration can apply to various substances in other contexts, in this ontology, it is specifically used for phosphoric acid. Finally, the "est Une Forme De" in English "is a form of" relationships between "Fluorure d'Hydrogène" (hydrofluoric acid) and "Acide Fluorosilicique" (fluor silicic acid) and hydrogen fluoride was introduced, illustrating the different forms in which these "Acide Fluorhydrique" (fluor acids) exist and their main component, hydrogen fluoride.

Combining these relationships and classes in the OWL ontology created a representative model of the links between the different chemical components. This provides a basis for a semantic representation of OCP Glossary items and their properties in a computational context, thereby facilitating better understanding and management of this chemical data. In order to test the performance of the proposed method, both the ONTO6 methodology [19] and the method cited in the article [17] were applied. The proposed method for ontology extraction from a glossary is distinguished by its hybrid approach, combining heuristic and algorithmic techniques. This method relies on both domain expertise and natural language processing techniques, including Jaccard similarity, to detect and structure relationships between glossary terms. Unlike more traditional methods, such as those mentioned in "into6.pdf", which rely mainly on semi-automatic techniques for extracting concepts and relationships from glossaries, this approach places particular emphasis on the accuracy and relevance of the generated ontologies, taking into account the linguistic and conceptual specificities of the domain studied.

Furthermore, while the ONTO6 methodology focuses on creating lightweight ontologies for the software testing domain, the proposed method aims to capture more complex and implicit relationships specific to the phosphate industry, making it more suitable for the needs of a highly specialized domain. Compared to the method [17], which relies on a series of iterations to refine the ontology, this approach favours a thorough analysis from the start, thus allowing the creation of a more robust ontology that is immediately usable in knowledge management systems.

After applying these two methods to a glossary of the same company, the glossary consists of 432. The following table shows a comparison of the results obtained. The results presented in the table provide a detailed comparison between the proposed ontology extraction method, the ONTO6 method, and the Evolutive Process, all applied to the same OCP glossary. The proposed method generated a higher number of properties (2826) and detected more relationships (85), reflecting its ability to capture complex relationships and provide a richer, more detailed ontology.

It also stands out in terms of accuracy with a 95% accuracy rate, significantly higher than that of the ONTO6 method (35%) and the Evolutive Process (72%). This highlights the effectiveness of the hybrid approach in accurately capturing domain-specific concepts and relationships. Regarding ontology structure, the proposed method produces a "Heavyweight" ontology well-suited to the specialized needs of the phosphate industry, in contrast to ONTO6, which generates a lighter ontology more suitable for less specialized domains. The Evolutive Process, meanwhile, produces a "Mediumweight" ontology, balancing between detail and generality.

Table 1. Comparative table of the results obtained

Comparison Point	The proposed method	ONTO6 Method [19]	Evolutive Process [17]
Number of classes	432	453	465
Number of Properties	2826	1785	2189
Relationship Detection	85	35	61
Accuracy	95%	35%	72%
Ontology Structure	Heavyweight	Lightweight	Mediumweight

5. Conclusion

Creating a tailor-made ontology adapted to the specific needs of the OCP Group in the phosphates industry marks a significant step forward in information management within the organization. This ontology represents a valuable resource, precisely capturing and structuring the specialized knowledge accumulated over the years. The integration of this ontology into an internal search engine opens the way to many promising perspectives. Indeed, it will improve the efficiency and relevance of search results, thus facilitating rapid access to crucial information for decision-making, problem-solving, and the development of new strategies. Our next task will involve developing a hybrid "semantic" search engine by combining this ontology with statistical methods.

This engine will utilize structured information from ontology to understand the meaning and relationships between concepts while leveraging statistical techniques to analyze trends and patterns in unstructured data. This hybrid approach will ensure even more accurate and relevant search results, providing an improved user experience and considerable added value for the business. Company employees will benefit from more efficient and targeted searches, reducing the time spent finding relevant information and increasing overall productivity.

Additionally, by leveraging the relationships and concepts defined in the ontology, the internal search engine can provide intelligent suggestions and personalized recommendations, thus promoting closer collaboration between different departments and teams within the company. This approach will also enable better knowledge and know-how management within the company, facilitating the sharing and transmission of knowledge between employees and promoting innovation and continuous development.

Additionally, by providing smoother onboarding for new employees, the internal search engine will help reduce adaptation times and accelerate their contribution to company projects and goals. In summary, integrating this ontology into an internal search engine represents a crucial step towards the digital transformation of the OCP Group, strengthening its ability to fully exploit its information resources and remain at the forefront of innovation in its field of activity.

Acknowledgements

The authors would like to thank the OCP group for their valuable discussions. This work is sponsored by the project “Knowledge Management for Development in the Context of OCP Group (KM4Dev – OCP Group)” granted by OCP Group, Morocco.

References

[1] Benamar Bouougada et al., “Mapping Relational Database to Owl Ontology Based on MDE Settings,” *Artificial Intelligence Review*, vol. 35, pp. 217-222, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[2] Batool Lakzaei, and Mehrnoush Shamsfard, “Ontology Learning from Relational Databases,” *Information Sciences*, vol. 577 pp. 280-297, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[3] Bilal Ben Mahria, Ilham Chaker, and Azeddine Zahi, “A Novel Approach for Learning Ontology from Relational Database: from The Construction to the Evaluation,” *Journal of Big Data*, vol. 8, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[4] Len Yabloko, The OntoBase-Protégé, 2023. [Online]. Available: https://protegewiki.stanford.edu/wiki/OntoBase

[5] Olivier Curé, Myriam Lamolle, and Chan Le Duc, “Ontology Based Data Integration Over Document and Column Family Oriented NOSQL,” *Arxiv*, pp. 1-16, 2013, [CrossRef] [Google Scholar] [Publisher Link]

[6] V.K. Kiran, and R. Vijayakumar, “Ontology Based Data Integration of NoSQL Datastores,” *2014 9th International Conference on Industrial and Information Systems (ICIIS)*, Gwalior, India, pp. 1-6, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[7] Hanen Abbes, Soumaya Boukettaya, and Faiez Gargouri, “Learning Ontology from Big Data through MongoDB Database,” *2015 IEEE/ACS 12th International Conference of Computer Systems and Applications (AICCSA)*, Marrakech, Morocco, pp. 1-7, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[8] Simin Jabbari, and Kilian Stoffel, “Ontology Extraction from MongoDB using Formal Concept Analysis,” *2017 2nd International Conference on Knowledge Engineering and Applications (ICKEA)*, London, UK, pp. 178-182, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[9] Bernhard Ganter, and Rudolf Wille, *Formal Concept Analysis: Mathematical Foundations*, Berlin Heidelberg: Springer-Verlag, 1st ed., pp. 1-370, 1999. [CrossRef] [Google Scholar] [Publisher Link]

[10] Yuangang Yao et al., “An Automatic Semantic Extraction Method for Web Data Interchange,” *2014 6th International Conference on Computer Science and Information Technology (CSIT)*, Amman, Jordan, pp. 148-152, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[11] Thomas Lampoltshammer, and Thomas Heistracher, “Ontology Evaluation with Protégé using OWLET,” *Infocommunications Journal*, vol. 6, no. 2, pp. 12-17, 2014. [Google Scholar] [Publisher Link]

[12] Gui-hyun Baek, Su-kyoung Kim, and Ki-hong Ahn, “Framework for Automatically Construct Ontology Knowledge Base From Semi-Structured Datasets,” *2015 10th International Conference for Internet Technology and Secured Transactions (ICITST)*, London, UK, pp. 152-157, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[13] Martin Seidel et al., "KESeDa: Knowledge Extraction from Heterogeneous Semi-Structured Data Sources," *Proceedings of the 12th International Conference on Semantic Systems*, New York, USA, pp. 129-136, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[14] Roberto Navigli, and Paola Velardi, "From Glossaries to Ontologies: Extracting Semantic Structure from Textual Definitions," *Proceedings of the 2008 Conference on Ontology Learning and Population: Bridging the Gap between Text and Knowledge*, pp. 71-87, 2008. [Google Scholar] [Publisher Link]

[15] Katarina Grolinger, Kevin P. Brown, and Miriam A.M. Capretz, "From Glossaries to Ontologies: Disaster Management Domain," *SEKE 2011 - Proceedings of the 23rd International Conference on Software Engineering and Knowledge Engineering*, pp. 402-407, 2011. [Google Scholar] [Publisher Link]

[16] Guntis Arnicans, Dainis Romans, and Uldis Straujums, "Semi-automatic Generation of a Software Testing Lightweight Ontology from a Glossary Based on the ONTO6 Methodology," *Frontiers in Artificial Intelligence and Applications*, vol. 249 pp. 263-276, 2013. [CrossRef] [Google Scholar] [Publisher Link]

[17] José R. Hilera et al., "An Evolutive Process to Convert Glossaries into Ontologies," *Information Technology and Libraries*, vol. 29, pp. 195-204, 2010. [CrossRef] [Google Scholar] [Publisher Link]

[18] Karin Koogan Breitman, and Julio Cesar Sampaio do Prado Leite, "Lexicon Based Ontology Construction," *Lecture Notes in Computer Science, Software Engineering for Multi-Agent Systems II Springer Berlin Heidelberg*, vol. 2940, pp. 19-34, 2004. [CrossRef] [Google Scholar] [Publisher Link]

[19] Loris Bozzato, Mauro Ferrari, and Alberto Trombetta, "Building A Domain Ontology from Glossaries: A General Methodology," *CEUR Workshop Proceedings*, Rome, Italy, vol. 426, pp. 1-10, 2008. [Google Scholar] [Publisher Link]

[20] Guntis Arnicans, and Uldis Straujums, "Transformation of the Software Testing Glossary into A Browsable Concept Map," *Lecture Notes in Electrical Engineering, Innovations and Advances in Computing, Informatics, Systems Sciences, Networking and Engineering*, vol. 313, pp. 349-356, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[21] Xianming Tang et al., "Construction and Application of an Ontology-Based Domain-Specific Knowledge Graph for Petroleum Exploration and Development," *Geoscience Frontiers*, vol. 14, no. 5, pp. 1-11, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[22] Tom Gruber, Ontology, Encyclopedia of Database Systems, Springer-Verlag, 2009. [Online]. Available: https://tomgruber.org/writing/definition-of-ontology/

[23] Bill Swartout et al., "Towards Distributed Use of Large-Scale Ontologies," *Association for the Advancement of Artificial Intelligence Spring Symposium Series on Ontological Engineering*, Stanford University, CA, pp. 138-148, 1997. [Google Scholar] [Publisher Link]