# Ontology-based NLP information extraction to enrich nanomaterial environmental exposure database

Ali Ayadi, Melanie Auffan, Jérôme Rose

Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 176 (2020) 360–369

Procedia Computer Science

www.elsevier.com/locate/procedia

24rd International Conference on Knowledge-Based and Intelligent Information & Engineering Systems

# Ontology-based NLP information extraction to enrich nanomaterial environmental exposure database

Ali Ayadi[a,*], Mélanie Auffan[a,b], Jérôme Rose[a,b]

[a]*CEREGE, CNRS, Aix-Marseille University, IRD, INRAE, Coll France, UMR 7330, 13545, Aix en Provence, France*
[b]*Duke University, CEE department, Durham, North Carolina, 27707, USA*

**Abstract**

In recent years, nanotechnologies have led to undeniable progress in any domains, such as electronics, materials and medicine. Despite the benefits of such a technology, a careful assessment of the potential risks for Human and Environmental health have to be studied. Assessing exposure and hazard to nanomaterials is a major challenge in the field of environmental sciences. This task requires to gather a large amount of meaningful experimental data usually generated by laboratory experiments. A first database of environmental exposure to nanomaterials (EXPOSED database) has been developed to gather data generated during mesocosm experiments. The challenge is now to enrich this database with more data from scientific articles in related fields. Herein, we present an ontology-based Natural Language Processing (NLP) approach to automatically extract and transfer data from text sources to database. This approach combines the use of NLP techniques and a domain ontology to automatically extract environmental exposure and hazards information. This approach was tested to enrich the EXPOSED database and indicators of quality highlight that this approach is effective and promising.



*Keywords:* Database enrichment; Information extraction; NLP techniques; Domain ontology; Engineered nanomaterials; Mesocosms.

## 1. Introduction

Nanotechnology is a multidisciplinary field of research involving among all physics, chemistry and biology, based on the exploitation of material properties at the atomic and molecular levels [1, 2]. It covers all the technologies used to manufacture, observe and measure objects, structures and systems with a size ranging from 1 to 100 nm [3]. In nanotechnology, engineered nanomaterials (ENMs) are intentionally manufactured by human activity for specific objectives [4]. These ENMs have gained a great deal of industrial interest because of their capacity to downsize devices and enhance properties of materials in multiple industrial sectors, such as electronics (by reducing the size of electronic

* Corresponding author. Tel.: +33 6 56 76 34 46.
*E-mail address:* ayadi@cerege.fr

1877-0509 
10.1016/j.procs.2020.08.037

components), health and cosmetics (new diagnostic tools), energy (new photovoltaic cells), etc. However, because of the presence of these ENMs in multiple objects of our daily, scientists must ensure about both their Environment and Human safety (called nanosafety) [5].

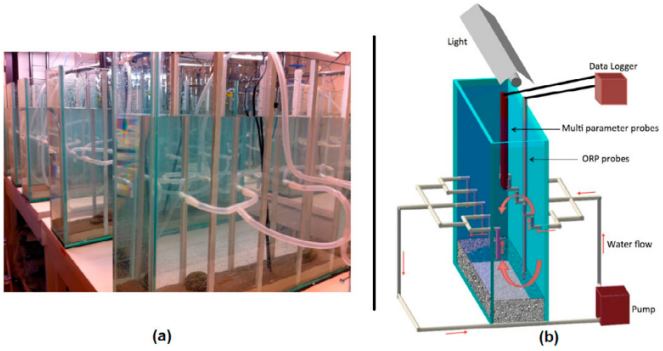


Fig. 1. Picture (A) and scheme (B) of 60L an indoor aquatic mesocosms platform [7].

Among all the testing strategies to study the environmental risk related to ENMs, mesocosms are particularly well suited (Figure 1) since they allow researchers to evaluate the exposure of ecosystems to ENMs and their hazards in environmentally relevant conditions. Mesocosms recreate indoor or outdoor an ecosystem, and measured different exposure, environmental, and hazard endpoints in response to ENMs contamination [6, 7, 8]. To date, experimental results acquired in mesocosms are grouped and stored in databases as the EXPOSED database[1] (for Environmental eXPOSure to Nanomaterials Database) developed by the SERENADE consortium (Safe(r) Ecodesign Research and Education applied to NAnomaterial DEvelopment). To date, these databases contained data that are directly collected during the experiments by the researchers and are manually enriched using textual documents (typically scientific articles and expert reports). Currently, this extraction of relevant information (called also ”data curation”) is fastidious, expensive and time consuming. One of the challenges in nanoinformatics is to develop automatic extraction of information for enriching nanosafety databases.

Several studies have focused on the exploration and use of text mining techniques in the field of nanotechnology. We distinguish four kinds of approaches, e.g. lexicon-based, rule-based, machine learning, and hybrid approaches. Firstly, *Lexicon-based approaches* use a lexicon, or dictionary of terms, to identify specific terms in the text. The main drawback of such approaches is that they cannot identify new entities in the nanotechnology domain. Moreover, the annotated corpus in each domain is done by experts from the domain and it requires time and effort to produce it. Examples include the works of Xiao L. et al. [10] who propose an entity recognition system based on a domain dictionary to extract chemical components without an explicit tokenization step. Secondly, *Rule-based approaches* involve designing approaches based on handwritten expert rules to identify relevant information. It requires a lot of manual work and is usually not reusable. In this category, we notice the works of Ykowiecka et al. [11] who describe a rule-based system for clinical data processing, and Akassi et al. [12] who propose a rule based tokenizer which uses manually extracted rule to identify chemical components. Then, *Machine learning approaches* use statistical models focused on recognizing specific entity names. The main disadvantage of these approaches is the dependence on annotated documents, which are difficult and expensive to obtain. Jessop et al. [13] propose, OSCAR4, an entity recognition system based on a Maximum-entropy Markov model, or Rocktäschel et al. [14] who propose a system based on conditional random field model to recognize chemical entities. Finally, *Hybrid approaches* combine different techniques to optimize the information extraction process. Examples include the works of García-Remesa M. et al. [15] who propose an approach for identifying specific nanotoxicity information using a Named Entity Recognition technique. Based on the natural language processing techniques, this method consists in recognizing named entities in a corpus and assigning them a label. Their proposed approach allows to identify instances of entities belonging to four

[1] https://aliayadi.github.io/EXPOSED-datbased-abase/#

categories: nanoparticles, exposure, toxicological, and targets. Jones D.E. et al. [16] propose an information extraction system based on natural language processing techniques for extracting numeric values of biomedical property terms of polyamidoamine dendrimers from nanomedicine literature. In their work, the authors integrate a domain ontology, the NanoParticle ontology, with the General Architecture for Text Engineering (GATE) and its information extraction module ANNIE [17].

In this paper, we demonstrate the use of natural language processing techniques and a domain ontology to extract environmental exposure to ENMs information from text sources, classify them according to the ontology concepts, and automatically enrich the database. From one hand, the NLP techniques are used to intelligently analyze textual documents and identify pertinent information from document [9]. On the other hand, the role of the domain ontology is to semantically classify and categorize the extracted information according to its concepts, thus associating the information with their corresponding attributes within the database. As a domain ontology, we use the EXPOSEO ontology[2], for environmental exposure to engineered nanomaterials ontology. This ontology was especially developed, with the assistance of experts, for modelling the domain knowledge of exposure-driven environmental risk assessment of ENMs. Herein, we used the EXPOSED database as a case study to test the feasibility of such an approach. We will evaluate using typical NLP metrics of recall, precision, and f-measure score whether the proposed approach can correctly extract exposure and hazard information from text sources and associate them with their corresponding attributes values in the EXPOSED database.

## 2. Proposed methodology

### 2.1. Architecture of the proposed methodology

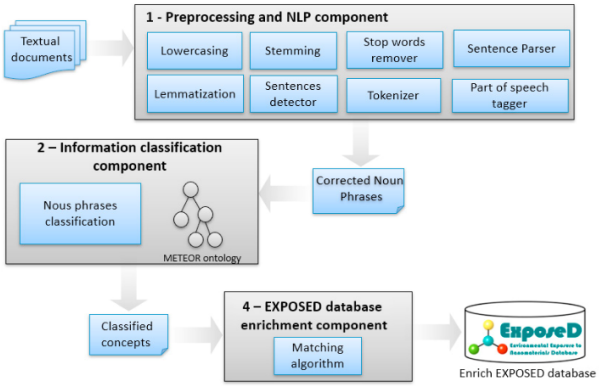


Fig. 2. Architecture of the proposed methodology.

Figure 2 illustrates the architecture of the proposed ontology-based NLP approach to automatically extract and transfer exposure and hazard data from text sources to the EXPOSED database, which comprises the following main components: a preprocessing and NLP component, an information extraction and classification component, and a database enrichment component. Detailed description of each component will be explained using a fictional running example involving some text lines taken from the scientific article of Tella M. et al. [6] as is shown in Figure 3.

### 2.2. Preprocessing and Natural Language Processing (NLP) component

This first component aims to transform raw data into an understandable format. This preprocessing component involves different tasks that are often complementary. It is an essential step to ensure the quality of the textual data

[2] https://github.com/AliAyadi/EXPOSEO_ontology

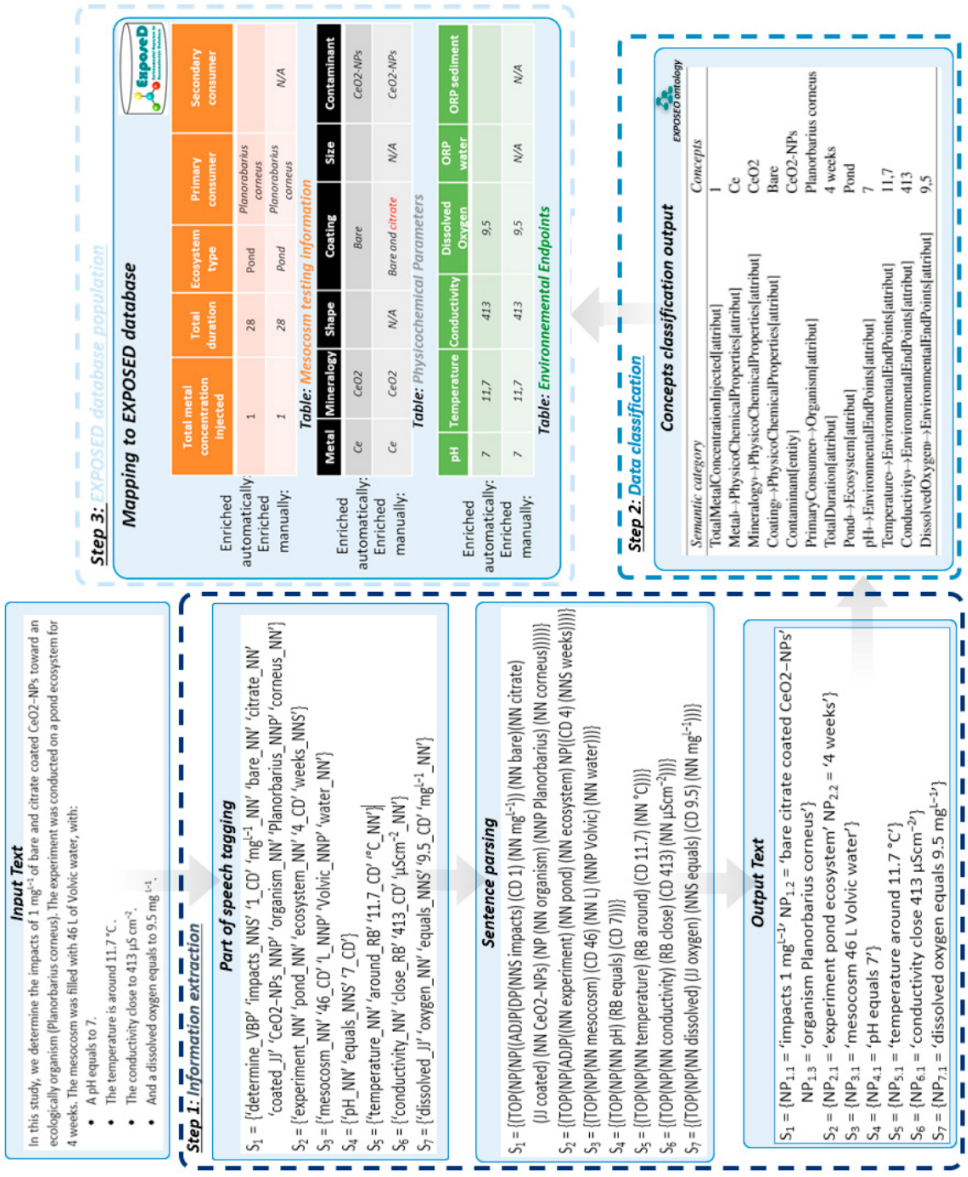


Fig. 3. An illustrative example showing the main steps of our proposed approach applied to some lines obtained from article [6].

to be analyzed. This component is mainly based on the Apache OpenNLP toolkit[3] which is a machine learning based toolkit for the processing of natural language text. In this step, we use different text preprocessing services, such as:

- **Normalization**: It consists in converting the text to the same case (in our case it is all lowercase), and removing the punctuation.
- **Stemming and lemmatization**: It allows representing under the same word a family of derivationally related words with similar meanings. This task cuts off the end or the beginning of the word, considering a list of

[3] https://opennlp.apache.org/

common prefixes and suffixes that can be found in an inflected word (e.g. ”affects”, ”affected”, ”affecting”, and ”affections” will become ”affect”). **Lemmatization** This task is less radical than the stemming. It retains the semantic meaning of the word but eliminates the gender or plural (e.g. ”nanoparticles” will be ”nanoparticle”).

- **Sentence boundary detection**: This process identifies where sentences begin and end. In our case, we use a rule-based sentence boundary detection open source method, the Python Sentence Boundary Disambiguation[4] (pySBD). Due to the particularity of our domain, we have used this sentence segmenter which extracts reasonable sentences when the format and domain of the input text are unknown.
- **Removing stop words**: This task consists of removing words belonging to stop words such as determiners, prepositions, etc. These are lists of words that have been previously defined from existing NLP libraries.
- **Tokenization**: It is a process dividing a string of characters into tokens (words), i.e. atomic elements of the string. The NLTK toolkit offers tokenizers already trained on a set of documents (or corpus). In our case, we only used the word tokenizer, especially the Penn Treebank word tokenizer.
- **Part-of-speech tagging**: It consists in identifying for each word its morphosyntactic class based on its context and lexical knowledge. We used the Tree tagger tool[5] which first allocates to the word its most used function and then infers another function using predefined rules. This task provides more semantics to the original text and enables us to get more granular information about its words.
- **Sentence parsing**: This task tries to define sub-components such as Name Entity and phrases in the sentence. It uses part-of-speech tagging as input and providing chunks as output. In our context, sentence parsing is very important because nanotoxicological entities are generally composed of words or noun phrases.

### *2.3. Information extraction and classification component*

After applying preprocessing techniques, we obtained a cleaner text that is significantly reduced in terms of size and format, which will increase the efficiency of the information extraction and classification task. This component ensures the task of recognition named entity and their classification. It aims at analyzing the nominal phrases provided by the previous component in order to identify and categorize relevant entities according to the semantic categories (names of ENMs, their physicochemical properties, mesocosms, the ecosystems they represent, etc.) predefined by a domain ontology. This will facilitate the matching of the extracted information with their corresponding attributes in the EXPOSED database, which is the role of the third component. As discussed in Section 1, the extraction and classification component is based on the EXPOSEO ontology. This domain ontology is used for modeling the domain knowledge of exposure-driven environmental risk assessment of engineered nanomaterials and provides a semantic knowledge base that will be used for the extracted noun phrases classification task, and associating them to their corresponding attributes in the EXPOSED database. The EXPOSEO ontology[6] was formalized in Web Ontology Language (OWL) using Protégé version 5.5.0. It has been evaluated and checked using the latest version of the Description Logic reasoner HermiT version 1.3.8. Even if we have used best-known validation methods to test the consistency of the EXPOSEO ontology, the intervention of domain experts is always necessary. In this paper, we referred to three experts who are respectively from computer science, chemistry, and biology to verify the experimental results manually. A cross-checking among them is made to ensure the authenticity of the assessment process. The domain experts evaluated the EXPOSEO ontology and concluded that it is in accordance with their knowledge in the environmental behavior, fate, and impacts of ENMs domain. Figure 4 shows an excerpt of the EXPOSEO ontology. The high level classes of the EXPOSEO ontology consists in the *Mesocosm experiment*, *Test_conditions*, *Physicochemical_Properties*, *Environmental_End_Points*, *Exposure_End_Points*, and *Hazard_End_Points* classes. It is also important to note that we have developed the EXPOSEO ontology following the ”Ontologies Inverse Engineering” technique [18], which consists of generating an ontology from a pre-existing database schema, in our case the EXPOSEO ontology was generated based on the EXPOSED database schema. This is mainly designed to facilitate the access to the information stored in the database. It allows to match the semantic categories provided by the EXPOSEO ontology used to categorize entities extracted from documents, with their corresponding tables and attributes in the EXPOSED

[4] https://spacy.io/universe/project/python-sentence-boundary-disambiguation
[5] https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/
[6] https://github.com/AliAyadi/EXPOSEO_ontology

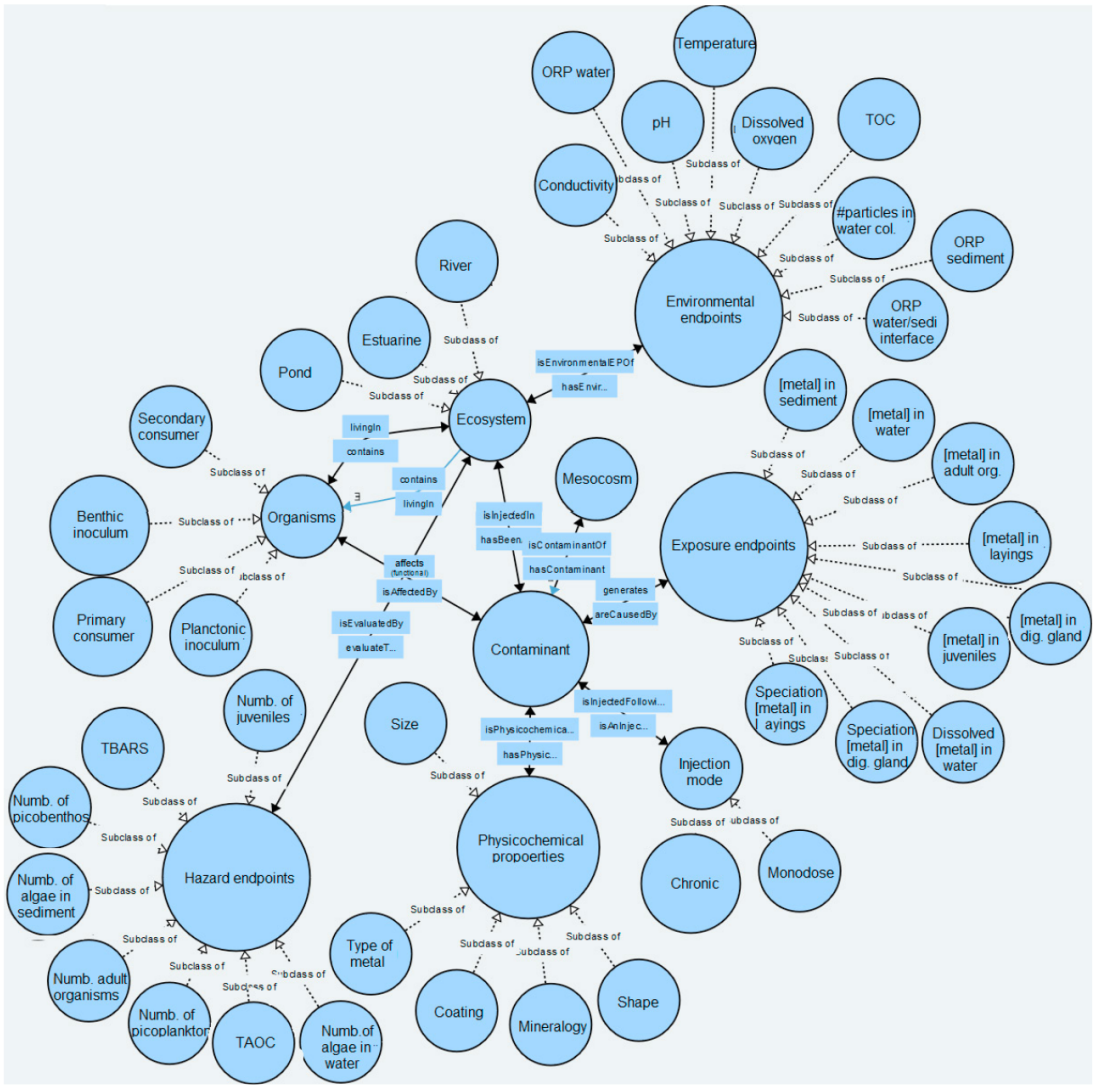


Fig. 4. An excerpt of the EXPOSEO ontology showing its higher-level classes.

database. This explains the fact that the EXPOSEO's classes and their subclasses (cited above) correspond to the different tables of the EXPOSED database and their attributes. That is why, in our approach we use the EXPOSEO ontology for recognizing and classifying the named entities of the extracted noun phrases. Firstly, a predefined dictionary has been developed based on the concepts of the domain ontology to ensure the recognition process. Indeed, the ontology concepts are structured in entities (class, subclass and instance) and relationships among them (object and data properties). The dictionary consists of the aggregation of all the ontology facts under common parent classes. To perform this task, we use the Jean-Baptiste Lamy's packages [19]. Using these packages and by querying the EXPOSEO ontology with the SPARQL language, we organized and classified the ontology content in a set of structured semantic categories (a dictionary). Thus, this dictionary contains all the collected data and grouped them into semantic groups including their class hierarchy and their type (class or attribute). For example, the group of concept *<TiO2, Ti, Ag, CuO, ...>* (representing respectively the titanium dioxide, titanium, silver, and cupper oxide) belongs to the semantic category *<Contaminant>* (representing the class *exposeo:Contaminant*). Another example, the group of value's attributes *<6nm, 10nm, 26nm, 30nm, ...>* belongs to the semantic category *<Size>* (representing the values

of the data property *exposeo:Size* of the class *exposeo:Contaminant*). Then, we use an algorithm, proposing by Engy et al. [20], for matching the entities composing the extracted noun phrases with their semantic categories provided by the dictionary. By the use of this algorithm and the predefined dictionary, this component can identify the extracted entities, and match them with their corresponding semantic category. The table representing the third step in Figure 3 shows some noun phrases associating with their semantic categories based on the EXPOSEO ontology.

The role of EXPOSEO ontology is to validate entities identified in the text by the automatic NLP techniques. Represented by description logics, it allows exploiting hierarchical relationships between concepts and implicit knowledge. Our domain ontology produces a very rich dictionary with more than 3900 axioms. According to Matentzoglu et al. that categorized ontologies based on their size [21], the EXPOSEO ontology is considered as a large ontology as it has more than 3900 axioms. The advantage of using this domain ontology results from its ability to infer new solutions through the subsumption relationship. For example, an entity that is typed *exposeo:River* (a river) is necessarily typed *exposeo:Ecosystem* (ecosystem). This ensures that the approach returns the most specific types to infer more generic types, and vice versa.

### 2.4. The database enrichment component

As described in the previous section, the generated semantic categories and the EXPOSED database schema are very similar. Indeed, the EXPOSEO's classes and their subclasses correspond to the different tables of the EXPOSED database and their attributes. Thus, the mappings here are simply the correspondences between each created ontological component (concept, property, etc.) and its original database component (table and attributes). This task is ensured by Algorithm 1 which takes as input both of the extracted entities with their corresponding semantic categories, and the schema of the EXPOSED database, then produces a mapping between elements of the two inputs that correspond to each other. Using Structured Query Language (SQL) and SPARQL Protocol and RDF Query Language (SPARQL) as data manipulation languages, this matching algorithm migrates the classified entities from their semantic categories to their appropriate positions in the EXPOSED database.

**Algorithm 1** Pseudocode of the EXPOSED database enrichment algorithm.

**Input:** List of extracted Noun Phrases ($NP_i \in N$), and their Semantic Categories ($SemanticCatergory(NP_i) \in S$).

**for** each $NP_i \in N$ **do**
  $Name\_Table = get\_Name\_Class(SemanticCatergory(NP_i))$
  $Name\_Attribut = get\_Name\_SubClass(SemanticCatergory(NP_i))$
  $Value\_Attribut = get\_Value\_SubClass(SemanticCatergory(NP_i))$
  $insert\_Database(Name\_Table, Name\_Attribut, Value\_Attribut)$
**end for**

**Output:** EXPOSED database enriched with extracted noun phrases.

## 3. Preliminary results

### 3.1. Dataset

We assembled a corpus of 10 scientific articles dealing with the exposure-driven environmental risk assessment of ENMs. These articles are available in the digital repository PubMed Central PMC database. Titles and references of these articles are available on Github[7]. These articles have been already curated manually by domain experts. From this 10 articles, we randomly extracted relevant sections (1345 entities) related to our ontology topic necessary. This corpus is used, firstly, for assessing the performance of our proposed approach to identify and classify relevant information in their appropriate location in the EXPOSED database, and secondly, as a gold standard to compare our obtained results with those obtained by the manually curation done by experts.

[7] https://github.com/AliAyadi/Articles

### 3.2. Experiments, results and analysis

After selecting the dataset, we implement our approach for identifying and classifying the candidate entities to enrich the EXPOSED database. The proposed approach was implemented in Python using the Natural Language Toolkit (NLTK), and the *Pandas*, *Matplotlib*, *Seaborn* and *Numpy* packages for performing the preprocessing and NLP tasks. We also use the Quest-Ontop platform [22] to ensure the mapping between SPARQL concepts and MySQL classes. From the domain ontology, we generated a dictionary representing the semantic categories with 20 categories (among 61 concepts), 45 different object properties (relations among concepts) and 56 different data properties (data properties representing the possible values of attributes), and more than 3000 unique instances. The EXPOSEO ontology and EXPOSED database are free to access in GitHub repositories.

A gold standard corpus (manually extracted reference standard corpus) is used to evaluate and compare the results obtained by our approach. Indeed, we compare our proposed method with the manually curation done by experts on the same corpus for the tasks of "Identification of relevant entities" and "Classification of entities" using the same corpus. Furthermore, these two tasks are analyzed for five different semantic categories. As shown in Table 1, we define five main categories corresponding to the main tables in the EXPOSED database and some of their attributes. Table 2 presents the obtained results of our approach (automatic approach) against the manually extracted reference standard using the different tables of the EXPOSED database. Functions **(1)** $Precision = \frac{Number\ of\ entity\ correctly\ identified\ /\ classified}{Number\ of\ entity\ identified\ /\ classified}$, **(2)** $Recall = \frac{Number\ of\ candidate\ entity\ correctly\ identified\ /\ classified}{Number\ of\ entity\ identified\ in\ the\ corpus}$ and **(3)** $F-measure = \frac{(2 * Precision * Recall)}{(Precision + Recall)}$ were used to compute each of Precision, Recall and F-measure respectively for evaluating the identification and classification tasks' performance in general (row "*All*" in Table 2), and also for the five main categories corresponding to the main tables in the EXPOSED database (rows $T_1$ to $T_2$ in Table 2).

The corpus consists of 1345 words. Among them, 1022 entities were identified manually with 996 correctly identified. Among these 996 entities, 984 entities have been classified with 970 entities correctly classified. These values correspond to a precision of 97.45%, a recall of 74.05% and an F-measure of 84.15%, for the task of identification of entities, and correspond to a precision of 98.57%, a recall of 97.38% and an F-measure of 97.97%, for the task of classification. Against the manual approach, our proposed method identifies a total of 1017 entities with 984 entities correctly identified. Among the correct identified entities, 971 entities have been classified with 952 entities correctly classified. These values correspond to a precision of 96.75%, a recall of 73.15% and an F-measure of 83.31%, for the task of identification of entities, and correspond to a precision of 98.04%, a recall of 96.74% and an F-measure of 97.38%, for the task of classification. These classified entities are themselves distributed according to the five semantic categories (or tables) previously defined as follows: (i) manually 156 entities for $T_1$, 247 entities for $T_2$, 225 entities for $T_3$, 178 entities for $T_4$, 150 entities for $T_5$, and (ii) automatically, 159 entities for $T_1$, 247 entities for $T_2$, 226 entities for $T_3$, 128 entities for $T_4$, 122 entities for $T_5$. For sake of space, we, only detail the results for the general process of identification and classification (row "*All*" in Table 2), however, detailed results of each category ($T_1$ to $T_5$) are presented in Table 2.

The proposed approach was compared with two similar approaches for entities extraction that use respectively the Nanotoxicological [10] and Nanoparticle [15] ontologies. The classification process is limited to five categories in Xiao et al. (material type, shape, capping agent, receptor, and particle size), and four categories in the García-Remesal et al. (nanoparticles, environmental exposure to nanoparticles, toxicological hazards of nanoparticles, and the targets of the hazards). Both lead to good precision, recall and F-measure (respectively of 82.90%, 79.70% and 81.30% in ref. [10], and 94.97%, 90.57% and 92.65% in ref. [15]). The Precision (98.04%), Recall (96.74%) and F-measure (97.38%) obtained with the EXPOSEO ontology-based NLP information extraction to enrich the EXPOSED database highlight the strength of our approach. Indeed, we reach similar evaluation metrics while using four times more categories.

The preliminary results presented in the current study highlight that the proposed approach is effective and promising in automatically identifying and extracting relevant entities from scientific articles. However, a thorough qualitative analysis of the results provides improvement pathways. Moreover, even if the tests on a small corpus are considerable, we will improve the identification task by adding a deep learning technique. It is also required to test our method using a larger dataset to enhance the performance and quality of the proposed approach. Indeed, using more peer reviewed documents regarding the environmental exposure and hazards of ENMs may affect the performance of the proposed approach. Finally, the implementation of a system prototype with sophisticated interfaces to simplify the interaction among the different components of the approach will be provide to users.

### 3.3. Error analysis

Manually, 26 terms were wrongly identified and 26 entities were incorrectly classified. With the proposed approach, 33 terms were unambiguously identified and 32 entities were incorrectly classified. It is obvious that the extraction process performs better in the first tables than in the last two tables. This is mainly due to some terms that the proposed approach succeeded in classifying them into more than one semantic category. While experts have classified them in only one category. Experts considered useful to maintain this classification since it is correct. This proves that, from a semantic point of view and synonymy of terms, the automatic approach is more efficient. Whereas in the other classes, which have distinct semantic categories, the automatic approach proves to be less efficient than the human expert. Moreover, most exposure and hazard endpoints are represented by compound entities, most have the same unit of measure, and some representing numbers of individuals without unit measure.

Table 1. The main tables of the EXPOSED database, used for the experiments, and some of their attributes.

| *Id* | *Table* | *Attributes* |
|---|---|---|
| $T_1$ | Initial conditions | (contaminant, total dose, total duration, ecosystem type, injection mode, organism) |
| $T_2$ | Physicochemical properties | (metal, mineralogy, shape, coating, size) |
| $T_3$ | Environmental endpoints | (ph, temperature, conductivity, dissolved oxygen, TOC, ORP water, ORP sediment) |
| $T_4$ | Exposure endpoints | ([metal] sediment, [metal] water, [metal] layings, [metal] adult, [metal] juveniles) |
| $T_5$ | Hazard exposure | (#picoplankton, #picobenthos, #algae water, #algae sediment, TBARS, TAOC) |

Table 2. The results of our proposed approach VS. the manual approach (P, R, F represents the precision, recall and F-mesure in %, respectively).

| Tables | Automatic approach | | | | | | Manual approach | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Identification | | | Classification | | | Identification | | | Classification | | |
| | *P* | *R* | *F* | *P* | *R* | *F* | *P* | *R* | *F* | *P* | *R* | *F* |
| $T_1$ | 90.85 | 88.83 | 89.82 | 96.36 | 90.85 | 93.52 | 86.66 | 84.32 | 85.47 | 94.54 | 89.14 | 91.76 |
| $T_2$ | 93.20 | 91.48 | 92.33 | 98.80 | 95.00 | 96.86 | 93.20 | 91.48 | 92.33 | 98.80 | 95.00 | 96.86 |
| $T_3$ | 90.00 | 88.23 | 89.10 | 95.74 | 91.83 | 93.74 | 90.00 | 88.23 | 89.10 | 95.74 | 91.83 | 93.74 |
| $T_4$ | 62.43 | 60.95 | 61.66 | 67.36 | 64.00 | 65.63 | 86.82 | 84.76 | 85.77 | 93.68 | 89.00 | 91.28 |
| $T_5$ | 69.71 | 67.77 | 68.72 | 76.25 | 71.76 | 73.93 | 85.71 | 83.33 | 84.50 | 93.75 | 88.23 | 90.99 |
| *All* | 96.75 | 73.15 | 83.31 | 98.04 | 96.74 | 97.38 | 97.45 | 74.05 | 84.15 | 98.57 | 97.38 | 97.97 |

Table 3. Performance comparison of the proposed approach and the state-of-the-art approaches in terms of precision, recall and F-mesure (in %).

| Xiao et al. approach [10] | | | García-Remesal et al. approach [15] | | | Proposed approach | | |
|---|---|---|---|---|---|---|---|---|
| *P* | *R* | *F* | *P* | *R* | *F* | *P* | *R* | *F* |
| 82.90 | 79.70 | 81.30 | 94.97 | 90.57 | 92.65 | 98.04 | 96.74 | 97.38 |

## 4. Conclusions and future work

This paper presented an approach for automatically extracting relevant information from scientific literature on the environmental risks associated to ENMs in order to enrich databases as EXPOSED. The proposed approach is considered as one of the first exploring the use of ontologies with NLP techniques in the field of nanotechnology. NLP techniques ensure extraction of knowledge from texts, while the EXPOSEO ontology ensures the classification of these extracted information. We also used a matching algorithm to associate the extracted information with their corresponding attributes in the EXPOSED database, and enrich it. The proposed method was tested in extracting ENMs exposure and hazard information from scientific articles using aquatic mesocosms already enter in the database. Obtained results were compared with the results of a fully manual method accomplished by the domain experts for each phase (identification, classification), in terms of precision, recall and F-measure. Apart database enrichment, this approach can be used to easily identify and classify relevant information for well-defined applications, such as discovering new knowledge for data analysis or comparison, enriching ontologies, etc.

## Acknowledgments

This work is a contribution to the Labex Serenade (No. ANR-11-LABX-0064) funded by the "Investissements d'Avenir" French Government program of the French National Research Agency (ANR) through the A*MIDEX project (No. ANR-11-IDEX-0001-02), and a contribution to the European project NanoInformaTIX, H2020-NMBP-TO-IND-2018-2020-814426. The authors acknowledge the Sustainable Environment Group of the CEREGE for helpful discussions and valuable feedbacks. This work is also a contribution to the OSU-Institut Pythéaas. The authors acknowledge the CNRS for the funding of the IRP iNOVE. The authors report no conflicts of interest.

## References

[1] Sahoo, S. K., S. Parveen, and J. J. Panda. (2007) "The present and future of nanotechnology in human health care." *Nanomedicine: Nanotechnology, Biology and Medicine*. **3.1**: 20-31.

[2] Mohamed, E. F., and Awad, G. (2019). "Nanotechnology and Nanobiotechnology for Environmental Remediation". *In Magnetic Nanostructures. Springer*: 77-93.

[3] Ricaud, M., and O. Witschger. (2012) "Les nanomatériaux." *Définitions, risques toxicologiques*.

[4] Lead, Jamie R., and Deborah M. Aruguete. (2010) "Manufactured nanoparticles in the environment." *Environmental Chemistry*. **7.1**: 1-2.

[5] Ray, P. C., Hongtao Yu, and Peter P. Fu. (2009) "Toxicity and environmental risks of nanomaterials: challenges and future needs". *Journal of Environmental Science and Health* **27.1**: 1-35.

[6] Tella, M., Auffan M., Brousset, L., Issartel, J., Kieffer, I., Pailles, C., Morel, E., Santaella, C., Angeletti, B., Artells, E., Rose, J., Thiéry, A., & Bottero, J. Y. (2014) "Transfer, transformation, and impacts of ceria nanomaterials in aquatic mesocosms simulating a pond ecosystem." *Environmental science & technology* **48.16**: 9004-9013.

[7] Auffan, M., Tella, M., Santaella, C., Brousset, L., Paillès, C., Barakat, M., Espinasse, B., Artells, E., Issartel, J., Masion, A., Rose, J., Wiesner, M., R., Achouak, W., Thiéry, A., & Bottero, J. Y. (2015) "An adaptable mesocosm platform for performing integrated assessments of nanomaterial risk in complex environmental systems". *Sci Rep*. **4**, 5608.

[8] Auffan, M., Masion, A., Mouneyrac, C., De Garidel-Thoron, C., Hendren, C. O., Thiery, A., Santaella, C., Giamberini L., Bottero, J., Wiesner, M., & Rose, J. (2019) "Contribution of mesocosm testing to a single-step and exposure-driven environmental risk assessment of engineered nanomaterials." *NanoImpact* **13**: 66-69.

[9] Russell-Rose, Tony, and Mark Stevenson. (2009) "The role of natural language processing in information retrieval: Searching for meaning and structure." *Information retrieval, searching in the 21st century*. 215-227.

[10] Xiao, L., Tang, K., Liu, X., Yang, H., Chen, Z., & Xu, R. (2013) "Information extraction from nanotoxicity related publications." *IEEE International Conference on Bioinformatics and Biomedicine. IEEE.*

[11] Ykowiecka, A., Marciniak, M., and Kupść, A. (2009) "Rule-based information extraction from patients' clinical data". *Journal of biomedical informatics*, **42(5)**, 923-936.

[12] Akkasi, A., Varoğlu, E., and Dimililer, N. (2016). "Chemtok: a new rule based tokenizer for chemical named entity recognition". *BioMed*.

[13] Jessop, D. M., Adams, S. E., Willighagen, E. L., Hawizy, L., and Murray-Rust, P. (2011). "OSCAR4: a flexible architecture for chemical text-mining". *Journal of cheminformatics*. **3(1)**, 1-12.

[14] Rocktäschel, T., Weidlich, M., and Leser, U. (2012). "ChemSpot: a hybrid system for chemical named entity recognition". *Bioinformatics* **28(12)**, 1633-1640.

[15] García-Remesal, M., García-Ruiz, A., Pérez-Rey, D., De La Iglesia, D., & Maojo, V. (2013) "Using nanoinformatics methods for automatically identifying relevant nanotoxicology entities from the literature." *BioMed research international*.

[16] Jones, David E., et al. (2014) "Automatic extraction of nanoparticle properties using natural language processing: NanoSifter an application to acquire PAMAM dendrimer properties." *PLoS One* **9.1**.

[17] Cunningham, H., Maynard, D., and Bontcheva, K. (2011) "Text processing with gate". *Gateway Press CA*.

[18] Abbasi, A. A., and Kulathuramaiyer, N. (2016) "A systematic mapping study of database resources to ontology via reverse engineering". *Asian Journal of Information Technology*, **15(4)**: 730-737.

[19] Lamy, J. B. (2017). "Owlready: Ontology-oriented programming in Python with automatic classification and high level constructs for biomedical ontologies". *Artificial intelligence in medicine*, **80**: 11-28.

[20] Yehia, E., Boshnak, H., AbdelGaber, S., Abdo, A., and Elzanfaly, D. S. (2019). "Ontology-based clinical information extraction from physician's free-text notes". *Journal of biomedical informatics*, **98**, 103276.

[21] Matentzoglu, N., Parsia, B., and Sattler, U. (2018). "OWL reasoning: Subsumption test hardness and modularity". *Journal of automated reasoning*, **60(4)**: 385-419.

[22] Rodriguez-Muro, M., and Calvanese, D. (2012)." Quest, an OWL 2 QL reasoner for ontology-based data access". *RodriguezCEUR-WS. org.*.