A SEMANTIC RETRIEVAL SYSTEM IN REMOTE SENSING WEB PLATFORMS

G.-A. Nys[1], J.-P. Kasprzyk[1], P. Hallot[2], R. Billen[1]

[1] Geomatics Unit, Department of Geography, ULiège - Place du 20 Août, 4000 Liège (ganys/JP.Kasprzyk/rbillen)@uliege.be
[2] LNA-DIVA, Faculty of Architecture, ULiège - Boulevard de la Constitution, 41 4020 Liège p.hallot@uliege.be

KEY WORDS: Ontology engineering, Remote sensing, NoSQL database, Natural language processing, Semantic

ABSTRACT:

This paper proposes a solution to reduce the semantic gap between final users and data/processing providers in a web market place dedicated to remote sensing products. Nowadays, search engine are common tools on the Internet. Users are accustomed to use them and used to get tabular classification of provided answers. These smart agents are set up to answer basic questions using automatic pages redirection or chitchat. In this research, to ensure coherence between user's requests and platform answers, natural language processing algorithms and knowledge graphs are integrated within a web platform thanks to a NoSQL graph database connected to open thesauri and Geographic Information Systems (GIS). Therefore, the most pertinent services can be proposed based on input sentences including non-technical vocabulary but also geographical components (the user interface includes a text area and an interactive map). While processing chains and remote sensing ontologies were presented in one of our previous studies, this article focuses on natural languages algorithms and knowledge mining.

1. INTRODUCTION

Remote sensing is not a well-known discipline outside research and education centres. The use of remote sensing services is commonly restricted to highly skilled professions (Lillesand, Kiefer, & Chipman, 2015). Normal users are rarely inclined to be interested in the discipline as getting results may become very time-consuming or need specific education.

Nowadays, search engine are common tools on the Internet. Users are accustomed to use them and used to get tabular classification of provided answers. Many engines answer complex natural language queries. Hidden in these virtual assistants, natural language processing (NLP) algorithms try to answer the users' queries by providing links to webpages or generic answers.

We study the possibilities to create a dedicated framework to reduce the gap between users (who are usually not familiar with remote sensing lexical field) and remote sensing services providers. For this purpose, we created a self-learning knowledge graph that structures the concepts used in remote sensing related queries. Queries are preprocessed by NLP algorithms in order to structure the concepts and reduce the fuzziness brought by natural languages and multilingualism. The complete workflow is defined as semantic system able to retrieve remote sensing services.

In a previous paper, we presented the development of an application ontology for the structuring of remote sensing operations shared by different processing chains (Nys *et al*, 2018). The main idea was to decompose processing chains, i.e. remote sensing services, into elementary operations linking different types of data. This decomposition allows the management of a web market place dedicated to remote sensing and services providing.

The remainder of this paper is structured as follows. First, we develop the technical workflow that is used to processes users queries through NLP: lemmatisation, Part-of-Speech tagging, geographical entity recognition, etc. After that, considerations about query expansion and terms dispatching within different modules are discussed. The thesauri reconstruction algorithm is an important part of the paper so a specific section develops advantages and disadvantages of the method. Geographical content of users' queries management finishes the technical workflow explanations. An example illustrates the different steps all along the paper. Finally, conclusion and future works describe possibilities and remaining challenges.

2. STATE OF THE ART

Among information retrieval algorithms, the one developed in Rocchio (1971) is particularly used in remote sensing research as a support of scenes and for change interpretation (Ghazouani *et al,* 2018). Moreover, NLP for information retrieval is no new domain but still younger than Rocchio's work. Lewis & Jones (1993) presented NLP indexing as a new effective method, which could easily supplant techniques of this time. They introduced more actual results which are summarised by Hirschberg & Manning (2015). More recently, Young *et al* (2018) introduced Deep Learning techniques in NLP support.

Today, ontologies are often used for effective knowledge modelling and information retrieval (Arvor *et al*, 2019). However, most of existing approaches based on ontologies generate relational database queries. In a more database-centred view, query formulation made with direct specification and "on-the-fly" manipulation is still not supported. Users commonly have a lack of understanding of query languages such as SQL. Therefore, reinforcement learning and other artificial intelligence techniques are explored to automate query formulation (Zhong, *et al*, 2017).

Generally guided approaches use ontologies to structure the well-known domain vocabulary and limit the queries possibilities within the scope of a specific field (Klien, *et al*, 2006; Lutz & Klien, 2006). Such an approach avoids the complexity and heterogeneity brought by natural language queries. In addition, this is sometimes done in a local way on limited geocatalogues (Shvaiko, *et al*, 2010). Moreover, ontologies can also be used in knowledge discovery within the scope of geographical information management (Bogdanović, *et al*, 2015).

Complexity of queries writing is also a remaining challenge when it comes to ontology uses in knowledge discovery (Munir & Sheraz Anjum, 2018). Some proposition tried to consider textual queries instead of simple words matching between lists (Mauro *et al*, 2017). Nevertheless, these examples do not reflect human languages complexity and limit their proposition to terms matching. Contextualisation is the key to keep query consistency and correct entity recognition.

Regarding the implementation, place names and toponyms can be stored in a complex knowledge graph. A directory of place names and toponyms is called a gazetteer. Such a geographical database may handle multilingualism and offers a solution when it comes to define a place with different names in multiple languages (Laurini, 2017). The most important gazetteer, GeoNames, is part of the YAGO project (Rebele *et al*, 2016) and proposes approximately eleven millions of entities that are freely available. Even if some cross-dataset and cross-lingual issues remain, it is currently the most popular open database of toponyms, especially for Belgium (Ahlers, 2017).

According to the state of the art, natural language processing tools and ontologies may reduce the semantic gap between non-specialist users and data/processing providers regarding answers to spatio-semantic queries. Structured around an application ontology implemented in a triple store database, NLP algorithms may enhance the communication inside remote sensing market places.

Natural language is a difficult thing to structure because it naturally evolves with humans' interactions through repetition and use. Ontologies may provide here a dynamic structure able to evolve but also to manage multilingualism. Natural language modifications are often made without conscious planning or premeditation. Considerations upon these statements are developed and studied.

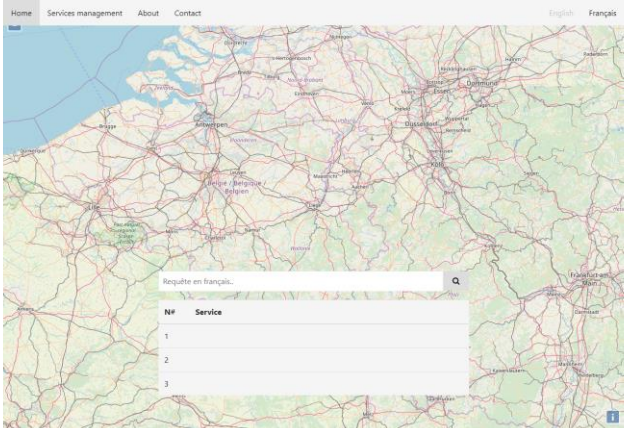
3. TECHNICAL WORKFLOW

3.1 Preliminary notes

The application ontology developed in one of our previous work (Nys et al., 2018) structures the processing chains proposition in a well-formalised knowledge graph. In this "Services Ontology", processing chains are defined within a specific class described following the Dublin Core metadata standard ontology (DC Terms): *dc:description* (http://purl.org/dc/elements/1.1/description). This "description" class is defined as followed: "*[…] an account of the resource. Description may include but is not limited to: an abstract, a table of contents, a graphical representation, or a free-text account of the resource*". In the scope of this research, description is applied to processing chains through natural language (e.g. "This service intends to recognise tree species").

The *countryTag* annotation property is another important property to notify. It defines the spatial coverage of the service. While certain services are directly impacted by the considered location, some may have a worldwide coverage. Indeed, it is trivial that processing chains on health status of the vegetation may be restricted to specific locations; vegetation is different in Africa than in Belgium. The country tags are based on ISO 3166-1 alpha-2 specifications. Note that a service with a worldwide coverage is tagged with "WW". This one was created in the scope of the project as an extension of the ISO proposition.

An illustration of the semantic retrieval system in a common internet browser is presented in Figure 1.



N°	Service
1	
2	
3	

Figure 1. Project web page - Semantic retrieval system

3.2 SKOS standard

Simple Knowledge Organisation System[1] (SKOS) is a W3C standard developing specifications to support the creation of thesauri, classification schemes … within the Semantic Web. Its interoperability is guaranteed by the ISO25964 (International standard for thesauri and interoperability with other vocabularies) and as it is necessary to structure natural languages databases in the scope of Open Linked Data. Multilingualism is easily handled within the standard.

Based on RDF/OWL DL vocabulary, SKOS standard presents well-defined relationships between entities and improves knowledge structuring within the graph. It is particularly suited for the design and management of natural language applications structured around graph mining and tree structure algorithms. In particular, the following relationships are used in our application (*skos:* is the predefined prefix of the SKOS vocabulary: http://www.w3.org/2004/02/skos/core#):

- *skos:prefLabel*: the preferred lexical label for a resource, an entity, in a given language. Its number is limited to one per concept;
- *skos:altLabel*: acronyms, abbreviations, spelling variants, and irregular plural/singular forms may be included among the alternative labels for a concept. Misspelled terms are normally included as hidden labels;
- *skos:broader*: relates a concept to a concept that is more general in meaning. It is the inverse relation of skos:narrower.
- *skos:related*: relates a concept to a concept through an associative semantic relationship;
- *skos:narrower*: relates a concept to a concept that is more specific in meaning. It is the inverse relation of *skos:broader*.

3.3 Global presentation

The workflow is illustrated in Figure **2**. Schema of the workflow as followed: blue diamonds are computation algorithms, red data silos are thesaurus and/or ontologies used within the scope of the project and green rectangle are intermediate or final data that are defined in next sections.

[1] https://www.w3.org/2009/08/skos-reference/skos.html

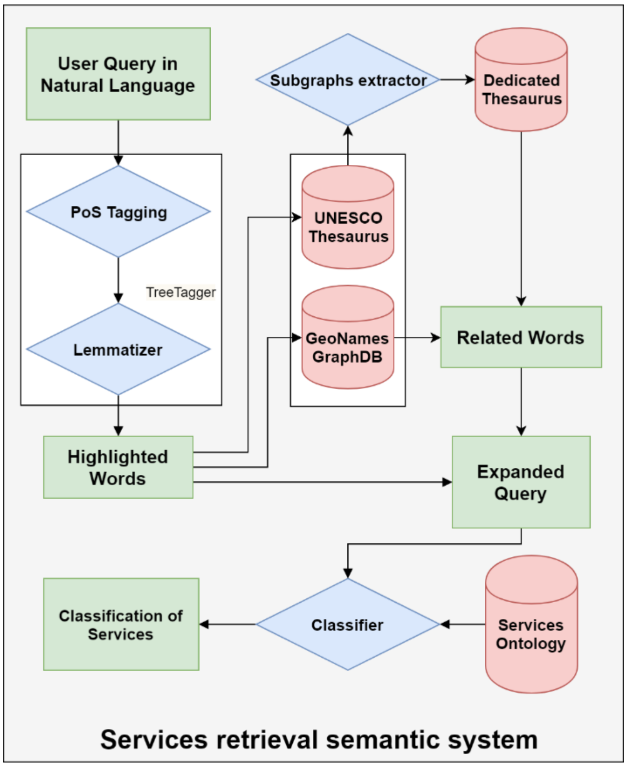


Figure 2. Schema of the workflow

For instance, the following natural language query, expressed in English, will illustrate each step through the semantic retrieval system (French is also supported in the current state of algorithms):

“How many trees are in a forest in Brussels?”

Note that this query is expressed in a relatively simple vocabulary but it is quite representative of what people commonly ask.

3.4 Natural Language Processing

NLP aims at teaching computers to understand and interpret human language by fractionating the elemental pieces of speech. It focuses on interactions between human languages and computers.

Computers are great at handling structured data such as relational tables or multidimensional arrays. However, human language is incredibly diverse and therefore not adapted to a rigid data structure. Some may be very complex. Human communication spans across thousands of languages and dialects including large sets of grammar rules, syntaxes and terms (especially French).

Therefore, NLP is a field that brings together computer science, artificial intelligence, big data and linguistics. Algorithms fractionate pieces of speech to understand natural language but they can also be used in the inverse way in order to mimic human language. Some answers can be found within this discipline especially with PoS Tagging and Lemmatization.

3.4.1 Part-of-Speech Tagging

Part-of-Speech tagging is the action of reading texts in some language and assigning parts of speech to each word depending on its role in the sentence (Noun, verb …). The PoS Tagger (piece of software that runs the algorithm) used in the scope of this research is TreeTagger[2], the Java version of the initial language-independent TreeTagger algorithm (Schmid, 1994, 1995). The main idea behind this work was to classify words through a decision tree trained on Penn-Treebank data (Marcus, Marcinkiewicz, & Santorini, 1993). Penn-Treebank classes list is the shortest classes list with 36 classes. We choose this list in a will to simplify the framework for the English part as state of the art stated. French part was trained on old French texts (Stein & Schmid, 1995). At the time, the classification provided better results than the well-known Trigrams (Cavnar & Trenlke, 1994) on the same data. Nowadays, it is still one of the most used and effective techniques.

Integrated in the project workflow, this part of the semantic retrieval system significantly impacts the computation time: at least 600ms are needed to calculate a piece of text, no matter how large it is, using the TreeTagger library. This point can be a problem when it comes to the production phase in a “user friendly” interface. Note that PoS Taggers do not correct any typo or grammatical mistakes. However, it manages full requests (e.g. “Where are the rice fields in Senegal?”) as well as terse requests (e.g. “Rice Field Senegal”).

Based on the section 3.3 example, words are classified by the PoS tagging with TreeTagger (trained on Penn-Treebank classes) as illustrated in Table 1. Example of PoS Tagging results:

Table 1. Example of PoS Tagging results

Word	Role	Word	Role
How	WRB – Wh-adverb	*a*	DT - Determiner
many	JJ - Adjective	*forest*	NN - Noun, singular or mass
trees	NNS - Noun, plural	*in*	IN - Preposition or subordinating conjunction
are	VBP - Verb, non-3rd person singular present	*Brussels*	NNP - Proper noun, singular
in	IN - Preposition or subordinating conjunction	*?*	SYM - Symbol

After the PoS tagging step, based on the computed tags, a filter is applied to extract the particular words that will influence the semantic content of the initial query. In particular, nouns, adverbs and verbs influence the intent hidden within the query. All the other tags (prepositions, symbols, etc.) will therefore be neglected in the following steps. The example can be reformulated as followed:

how[WRB] many[JJ] trees[NNS] are[VBP] forest[NN] Brussels[NNP]

3.4.2 Lemmatization

Lemmatization is the process that simplifies a word by removing the influence of secondary elements like conjugation, inflectional endings, etc. This aspect is particularly complex with the french language where gender (masculine/feminine) and number (singular/plural) of nouns both influence the spelling (and sometimes the pronunciation) of adjectives. Therefore, there is a need to simplify tagged words to reach a better understanding of the described concepts. Previous PoS Tagging step allows lemmatization in a consistent way. On the contrary running lemmatization first would not be appropriate since PoS Tagging

[2] https://reckart.github.io/tt4j/

needs context, which is translated in words inflected forms. This technique is different from Stemming which simply removes inflected forms from words and so get a stem. The same lemma can correspond to forms with different stems (e.g. verbs conjugation).

Finally, duplicates are deleted. Indeed, a service is not more useful if any words are used multiples times in its definition. Thereby, the PoS Tagging and Lemmatization algorithms process services descriptions in order to reduce their complexity and highlight their semantic potential. This deletion is made on both services description and users queries to keep consistency and avoid unscrupulous definitions that could skew results by repeating an important term many times.

Based on the results of the previous step, highlighted terms are:

how many tree be forest Brussels

3.5 Terms dispatching

After NLP algorithms, highlighted terms are dispatched in three different ways: some are exceptions that are not "expanded" (section 3.5.1) because of their conflictual nature, some are processed to extract knowledge from the reference thesaurus (section 3.5.2) and finally, some may add information about the spatial context of the query (section 3.5.4)

3.5.1 Query expansion

According to Grootjen & van der Weide (2006) knowledge can be extracted from a huge set of documents in a specific domain. However, such a semantic directory, a corpus, does not exist for remote sensing or related queries. Therefore, as it is not possible to train N-Grams algorithms (Damashek, 1995) or similar techniques, we decide to create a dedicated thesaurus as proposed in

Moreover, the thesaurus has to be structured following the SKOS standard, which greatly defines the relationships between concepts. This point is primordial for the following algorithms (section 3.5.2) while the choice of the source thesaurus is motivated based on its reliability. According to (Mandala *et al*, 1999), we restrict the number of source to one for performance of query expansion techniques: the UNESCO thesaurus.

The UNESCO thesaurus[3], created in 1977 and still under revision, structures and controls lists of terms in many fields: education, culture, natural sciences, social and human sciences, communication and information. Therefore, the following techniques are easily transposable in fields different from remote sensing. Moreover, the database is continuously enriched and updated through the different UNESCO's programmes and activities. This adds robustness for the algorithms but some missions and their domain can be neglected. The nature sciences part of the thesaurus is nevertheless sufficient in the scope of this project.

Behind the idea of extension in the Query Expansion, there is a need to limit the spread, in other words the dilution, of the original meaning of the query. For example in a more global context, replacing every word in a sentence by a synonym may bring fuzziness and mistakes in services classification.. This extension may go as far as to make the request irrelevant and therefore the answer too. This fuzziness could be established on a logical and mathematical basis (Buckley, Salton, & Allan, 1994).

In order to limit this phenomenon, on the one hand, some words are considered as exceptions and therefore are not processed by the following tools. For instance, the term "être" in French may be translated by the infinitive form of the verb "to be" or by the noun "being" (i.e. human being). These skip the "query expansion" step if PoS Tagging and Lemmatization did not provide a sufficient result. On the other hand, only the users' requests are expanded on the assumption that providers are precise enough in the description of services. Moreover, the expansion of both, users' queries and Services descriptions would lead to too much uncertainty. Note that services descriptions are however processed by NLP algorithms to reduce the complexity of their definitions.

3.5.2 Subgraph extraction

While SPARQL query language does support direct construct queries, which return a set of relations within a graph, the subgraph extraction here is made up of select queries to master each element. The algorithm of graph mining works sometimes through an API connected directly to the triple store, sometimes with a common SPARQL endpoint, depending on the reference thesaurus. Both query and storage strategies have their advantages and disadvantages but none is neglected in the scope of our research (Fernández *et al*, 2018).

Whenever a highlighted term matches a concept of the reference thesaurus, the subgraph of its nearest neighbours is extracted. Given that each concept is referred with a Unique Resource Identifier (URI), interactions and merging of different subgraphs are possible.

Besides other relations, the broader ones, explained earlier in this document, are extracted. This process runs until there is no broader relation and ends with a tree of concepts linked to top concepts. The top concept in Semantic Web is defined as "Thing". On the contrary, the other extremum is "Nothing". Everything is a "Thing" and no thing is "Nothing". It is one of the constituent Semantic Web hypothesises.

An example is illustrated in Figure **3**. Subgraph extraction where the red entity refers to the highlighted term. Starting from there, the algorithm traverses the graph through broader relations until the "Thing" concept is reached. During the graph traversal, linked concepts (narrower and related) are also included in the subgraph extraction. We limit the extraction to the first neighbours (first degree). Note that empty relations are represented here for further merging thanks to the use of URIs and the open world assumption.

[3] http://vocabularies.unesco.org/browser/thesaurus/en/

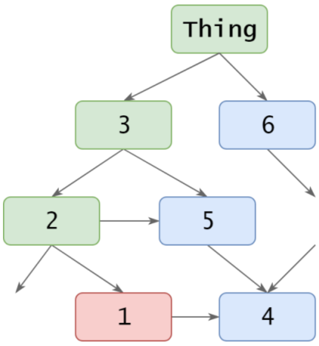


Figure 3. Subgraph extraction

During the successive iterations of the algorithm (pseudo-code in Table 2. Subgraph extraction algorithm), the dedicated graph expands and self-structures. The more the iterations, the more relevant the queries are.

Table 2. Subgraph extraction algorithm

```
function extract Subgraph(highlighted_terms)
  load dedicated_graph

    foreach concept in highlighted_terms do
      query surr_subgraph of concept in reference_graph

      forearch entity in surr_subgraph
        if entity is not in dedicated_graph
          add entity to dedicated_graph
            if entity is a vertex
              findBroader(entity)
            end if
        end if
      end foreach
    end foreach

  save dedicated_graph
end function

function findBroader(vertex)
  query broader, broader_relation of vertex in reference_graph

  if broader exist
    add broader to dedicated_graph
    add broader_relation to dedicated_graph
    findBroader(broader)

  end if
end function
```

One key to structure this new graph is to store the existence of a relation between the highlighted concepts and those we do not already know. This allows the reduction of the graph complexity while maintaining an anchor for the future graph fusions/additions. Remember that such anchors are mandatory because of the Open World Assumption and this is possible thanks to the use of the URIs. These URIs define every edges and vertices of the graphs. Relation of equivalence may exist between different graphs and these relations provide a way to merge third-party thesauri.

Table 3 shows an example of neighbours of the term “tree” as structured in the UNESCO thesaurus. Broader, related and narrower terms are all taken into account but do not in the same way as explained further.

Table 3. Example for tree[4] term in UNESCO thesaurus

Broader terms	Plants	Indeed, every tree is a plant. In linguistic, we call it a hypernym.
Related terms	Forest resources Forestry Forests Wood	These not only synonyms but also concepts, which are perceived as similar to the nodal word.
Narrower terms	Oak Poplar Fir	Here are species of trees. Another example could be Bread for Baked products.

3.5.3 Dedicated thesaurus reconstruction

Once the relevant information is extracted, there is a need to add it to the dedicated thesaurus in a consistent way. This step is part of a machine learning process to enhance the classification and the users’ query mining. The more the application will be used, the more accurate the classification will be. Consequently, the relevance of the database will increase with its uses. Indeed, a well-trained tool is the consequence of many queries.

As specified above, SKOS language, as a RDF/OWL DL vocabulary, allows an easy merging of different information sources, as long as they are well structured. This point is mandatory when it comes to combine newly extracted information within the current state of the knowledge base. In the reference thesauri, many parts of the databases could be irrelevant in the mentioned application. This is especially true with the UNESCO one where many sciences fields are studied but not related to remote sensing (Politics, economics…).

Therefore terms in users’ queries influence the data training so that the dedicated knowledge base is constituted of the most used and accurate terms. Nevertheless, overfitting with other fields is not considered since algorithms are suited for this particular application.

3.5.4 Geographical content

The geographical component of a query is a predominant aspect when it comes to remote sensing. Nevertheless, it is not relevant to manage it through a natural language thesaurus. Indeed, the spatial nature of the geographical component needs another method that considers spatial analysis concepts like distance, spatial entity, topology, coordinates reference system, analysis scale, etc. In order to deal with this geographical aspect, GeoNames proposes access to the biggest open geographical graph database, which contains more than eleven millions place names.

The management of the geographic content of a query is distinguished in several parts:

- Contextualisation based on toponyms and place names:
    - Nearest neighbours
    - Administrative subdivisions
- Contextualisation using the background map.

[4] http://vocabularies.unesco.org/thesaurus/concept2672

Indeed, there is a need to limit the query expansion, as it was done in the natural language thesauri. For instance, "Saint-Louis" is a well-known city in the north of Senegal, a city in Missouri (United States) and even a place in Belgium near Courtrai. In fact, toponymy may bring fuzziness if there is no additional information like positioning distance from a central position or bounding boxes. In this context, there is a need to provide such a positioning and it is easily done by clicking the map background as shown in Figure **1**. Project web page - Semantic retrieval system. About position, three filters are implemented to expand the query in a more relevant way:

First, mouse-clicking position in the background map is used to restrict services to a specific country. For this purpose, each service is tagged with the relevant countries (our application only concerns Belgium and Senegal) in which the service provides relevant results (refer to section 3.1). For instance, a tree species recognition service for Belgium could not be used in Senegal and conversely, because of the different environment.

Secondly, coordinates of the clicked point are taken into account to find the administrative subdivisions that concern the query: city, borough, district, region, NUTS classification, etc. These influence the classification of services just as broader terms do: region is the generalization of a city; a country is the generalisation of a region… These terms can be present in services descriptions and thus be considered as narrower terms. Highlighted proper nouns are also taken into account for this aspect just like positioning.

Finally, the terms highlighted by the PoS Tagging, are sent to the Geonames database. The new extracted terms are the nearest administrative entities of the initial term. The weight attributed to these new terms is the same as for the related terms from the thesaurus.

All the previous statements and their corresponding steps in the workflow can easily be neglected if geographical information is not given in the user's query: map not clicked or no place name in the sentence.

3.6 Services classification

The last step before returning queries results is the Services Classification. For each candidate service, different arrays of matching terms are computed by the classifier: one for the highlighted terms, one for the broader terms, one for the related terms and one for the narrower terms (including the corresponding geographically tagged terms). After that, the classifier sums up arrays occurrences in order to obtain a score for each service. The sum is weighted as indicated in Table **4** (currently, weights are determined empirically after tests with one hundred composition tables). The final output is a list of services sorted by their score.

Table 4 | Arrays weights for classifier

Highlighted terms	**Broader terms**	**Related terms**	**Narrower terms**
1	0.25	0.25	0.5

3.7 Interaction between ontologies

It is not strictly speaking a link between the two ontologies (the one dedicated to this project and the one structuring the services (Nys et al., 2018)) but rather an association. This association gets information from the reference thesaurus to enhance the semantic potential of data stored in the Services Ontology. Such a connexion is made on the fly and nothing remains of the modifications made by the processing in the former ontology. This in a process to leave both the ontologies independent of one another and therefore is considered as an association. People may choose to use each ontologies independently and therefore modularity is maintained.

It will then be possible to use ontologies and thesauri in different projects and applications, in the context of web market platform or not. Moreover, it is possible to take other combination of ontologies for scalability, languages changes or domain changes. The project is part of a dynamic that is increasingly focused on the pooling of knowledge: Semantic Web, Linked Data, Open Data … whatever it is called. There is a need to respect this condition for standardisation and accessibility. Some may find interest in other thesauri or ontologies and no possibilities are therefore neglected.

4. CONCLUSION AND FUTURE WORK

Natural language processing algorithms, thesauri and knowledge graphs may be used in support of semantic retrieval systems. The former is mandatory to allow terms recognition and highlighting terms. Working on processes and analysis on large amounts of natural language queries and/or products description allows reducing the semantic gap between machines and humans. This is also useful to reduce semantic gap on web platform between users that are not familiar with the domain and professionals.

The dedicated graph reconstruction proved its usefulness in supporting web applications. Semantic web technologies, like Simple Knowledge Organisation System, are mandatory to reach such a purpose. Algorithms were developed in a will to preserve scalability and modularity. Indeed, reference database, languages, thesauri … every step is modular following the purpose of the reconstructed thesaurus. We proved the usefulness of such an approach through users' usages on an open web platform.

Future work will study the scalability of such a system by integrating new languages and new reference thesauri. Note that libraries used in the scope of this project already support many languages and discourse domains. The merging of different sources is a great incoming challenge. Scalability of services number also needs to be studied.

ACKNOWLEDGEMENTS

This research is supported by the EORegions_Science project, which is funded by the Belgian Science Policy Office (BELSPO).

The authors would like to thank the partners of the EORegions project for their collaboration and information. Finally, the authors are grateful to the reviewers for their comments.

REFERENCES

Ahlers, D. 2017. Linkage Quality Analysis of GeoNames in the Semantic Web. *Proceedings of the 11th Workshop on Geographic Information Retrieval - GIR'17*, pp. 1–2. https://doi.org/10.1145/3155902.3155904

Arvor, D., Belgiu, M., Falomir, Z., Mougenot, I., & Durieux, L. 2019. Ontologies to interpret remote sensing images: why do we need them? *GIScience & Remote Sensing*, pp. 1–29. https://doi.org/10.1080/15481603.2019.1587890

Bogdanović, M., Stanimirović, A., & Stoimenov, L. 2015. Methodology for geospatial data source discovery in ontology-driven geo-information integration architectures. *Journal of Web Semantics*, *32*, pp. 1–15. https://doi.org/10.1016/j.websem.2015.01.002

Buckley, C., Salton, G., & Allan, J. 1994. The Effect of Adding Relevance Information in a Relevance Feedback Environment. *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Presented at the SIGIR94, Dublin, Ireland.

Cavnar, W. B., & Trenlke, J. M. 1994. N-Gram-Based Text Categorization. *Proceedings of the 3rd Annual Symposium on Document Analysis and Information Retrieval*. Presented at the SDAIR-94.

Damashek, M. 1995. Gauging Similarity with n-Grams: Language-Independent Categorization of Text. *Science*, *267*(5199), pp. 843–848. https://doi.org/10.1126/science.267.5199.843

Fernández, J. D., Umbrich, J., Polleres, A., & Knuth, M. 2018. Evaluating query and storage strategies for RDF archives. *Semantic Web*, pp. 1–45. https://doi.org/10.3233/SW-180309

Ghazouani, F., Farah, I. R., & Solaiman, B. 2018. Semantic Remote Sensing Scenes Interpretation and Change Interpretation. In C. Thomas (Ed.), *Ontology in Information Science*. https://doi.org/10.5772/intechopen.72730

Grootjen, F. A., & van der Weide, T. P. 2006. Conceptual query expansion. *Data & Knowledge Engineering*, *56*(2), pp. 174–193. https://doi.org/10.1016/j.datak.2005.03.006

Hirschberg, J., & Manning, C. D. 2015. Advances in natural language processing. *Science*, *349*(6245), pp. 261–266. https://doi.org/10.1126/science.aaa8685

Klien, E., Lutz, M., & Kuhn, W. 2006. Ontology-based discovery of geographic information services—An application in disaster management. *Computers, Environment and Urban Systems*, *30*(1), pp. 102–123. https://doi.org/10.1016/j.compenvurbsys.2005.04.002

Laurini, R. 2017. Gazetteers and Multilingualism. In *Geographic Knowledge Infrastructure* (pp. 157–182). https://doi.org/10.1016/B978-1-78548-243-4.50008-6

Lewis, D. D., & Jones, K. S. 1993. Natural language processing for information retrieval. *Communications of the ACM*, *39*(1), pp. 92–101. https://doi.org/10.1145/234173.234210

Lillesand, T. M., Kiefer, R. W., & Chipman, J. W. 2015. *Remote sensing and image interpretation* (Seventh edition). Hoboken, N.J: John Wiley & Sons, Inc.

Lutz, M., & Klien, E. 2006. Ontology-based retrieval of geographic information. *International Journal of Geographical Information Science*, *20*(3), pp. 233–260. https://doi.org/10.1080/13658810500287107

Mandala, R., Tokunaga, T., & Tanaka, H. 1999. Combining multiple evidence from different types of thesaurus for query expansion. *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '99*, pp. 191–197. https://doi.org/10.1145/312624.312677

Marcus, M., Marcinkiewicz, M. A., & Santorini, B. 1993. Building a large annotated corpus of English: the penn treebank. In *Computational Linguistics - Special issue on using large corpora: II* (Vol. 19, pp. 313–330). Cambridge, USA: MIT Press.
Mauro, N., Ardissono, L., & Savoca, A. 2017. Concept-aware geographic information retrieval. *Proceedings of the International Conference on Web Intelligence - WI '17*, pp. 34–41. https://doi.org/10.1145/3106426.3106498

Munir, K., & Sheraz Anjum, M. 2018. The use of ontologies for effective knowledge modelling and information retrieval. *Applied Computing and Informatics*, *14*(2), pp. 116–126. https://doi.org/10.1016/j.aci.2017.07.003

Nys, G.-A., Kasprzyk, J.-P., Hallot, P., & Billen, R. 2018. Towards an ontology for the structuring of remote sensing operations shared by different processing chains. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *XLII–4*, pp. 483–490. https://doi.org/10.5194/isprs-archives-XLII-4-483-2018

Rebele, T., Suchanek, F., Hoffart, J., Biega, J., Kuzey, E., & Weikum, G. 2016. YAGO: A Multilingual Knowledge Base from Wikipedia, Wordnet, and Geonames. In P. Groth, E. Simperl, A. Gray, M. Sabou, M. Krötzsch, F. Lecue, … Y. Gil (Eds.), *The Semantic Web – ISWC 2016* (Vol. 9982, pp. 177–185). https://doi.org/10.1007/978-3-319-46547-0_19

Rocchio, J. J. 1971. Relevance Feedback in Information Retrieval. *The SMART Retrieval System - Experiments in Automatic Document Processing*.

Schmid, H. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of the International Conference on New Methods in Language Processing*. Presented at the International Conference on New Methods in Language Processing, Manchester, United Kingdom.

Schmid, H. 1995. Improvements In Part-of-Speech Tagging With an Application To German. *Proceedings of the ACL SIGDAT-Workshop*, 47–50.

Shvaiko, P., Ivanyukovich, A., Vaccari, L., Maltese, V., & Farazi, F. 2010. A semantic geo-catalogue implementation for a regional SDI. *Proceedings of the INPIRE Conference 2010*. Presented at the INPIRE conference 2010, Krakow, Poland.

Stein, A., & Schmid, H. 1995. Etiquetage morphologique de textes français avec un arbre de décisions. *TAL. Traitement Automatique Des Langues*, *36*(1–2), 23–35.

Young, T., Hazarika, D., Poria, S., & Cambria, E. 2018. Recent Trends in Deep Learning Based Natural Language Processing. *IEEE Computational Intelligence Magazine*, *13*(3), 55–75. https://doi.org/10.1109/MCI.2018.2840738

Zhong, V., Xiong, C., & Socher, R. 2017. Seq2SQL: Generating Structured Queries from Natural Language using Reinforcement Learning. *ArXiv:1709.00103 [Cs]*. Retrieved from http://arxiv.org/abs/1709.00103