Semantic Filtering of Scientific Articles guided by a Domain Ontology

Fabrice Jouanot[1], Cyril Labbé[1], Elena Michael[1,2], Marie-Christine Rousset[1], Maithé Tauber[3], Federico Ulliana[1]

***Abstract* — The problem that we address in this paper is how to improve the accuracy of retrieving specialized information within a textual scientific corpus. We present a new approach in which the keywords expressing the bibliographical needs of a researcher are related to a domain ontology. We illustrate how such a declarative ontolology-based approach can be used both for computing varied statistics, and also for helping experts to find useful fine-grained information within a textual corpus.**

***Index Terms*— Concept extraction from text, information integration, ontology-based data access, semantic search.**

I. Introduction

For researchers, finding the bibliographical articles likely to improve their knowledge on their domain of expertise is a crucial but time-consuming task. Subscription keywords-based search tools exist to help researchers monitoring the web or bibliographic databases of their domain. However, due to the limitation of keywords to capture fine-grained knowledge, the bibliographical articles returned by keywords filtering need a further analysis to retain those that are really relevant to the needs of the researcher. Up to now, this analysis is done manually by the researcher herself of by a human assistant.

In this paper, we present a new approach in which the keywords expressing the bibliographical needs of a researcher are related to a fine-grained description of her domain of expertise in the form of an ontology. An ontology is a formal description providing human users a shared understanding of a given domain, and can also be interpreted and processed by machines thanks to a logical semantics that enables inference and reasoning.

In this paper, we first summarize the methodology underlying our approach, that takes as input a domain ontology and a corpus of articles, and constructs automatically a knowledge base enriching the input domain ontology with the relevant content and structure of the corpus. Then we illustrate through examples, the interest and the power of such a declarative ontology-based approach both for computing varied statistics on the corpus and its correlation with formal terms of a domain of interest, and also for helping experts to find useful fine-grained information within a textual corpus.

II. Methodology

Our approach consists in using a domain ontology for extracting the relevant content of a given bibliographical corpus and for representing it as a knowledge base equipped with reasoning and querying capabilities.

Step 1 is the design of ontology that can be either created manually by a domain expert or extracted from a pre-existing one. The ontology is provided using Semantic Web standards [2] (namely, RDFS enriched with rules), and stored as a set of RDF triples and a set of rules. Each formal term introduced by the expert is associated with a set of textual expressions that correspond to it in the scientific literature. This results in a table of correspondence used as input in the second step.

Step 2 consists in the automatic extraction from each paper in the corpus of the sentences that contain textual expressions corresponding to formal terms of the ontology. This is done using a suite of simple Perl scripts that take as input the corpus and the table of correspondence.

Step 3 consists in automatically assigning each paper and each extracted sentence to unique identifiers that are typed by the classes **Papers** and **Sentences** respectively. This information is encoded into RDF triples that are automatically inserted into the domain ontology. RDF triples expressing that each paper's identifier is related to the paper's title are also inserted in the ontology by using the RDF property **rdfs:label**. Similarly, each sentence identifier is associated to its textual content by using the RDF property **rdf:comment** and the corresponding triples are injected in the ontology. We have also defined the following properties for which the corresponding triples are also automatically injected in the ontology:

- **publishedIn**: to relate a paper's identifier to its year of publication,
- **isaSentenceOf**: to relate a sentence's identifier to the paper it belongs to
- **mentions**: to relate a sentence to a formal term (instance, class or property) of the domain ontology,
- **isAbout**: to relate a paper to a formal term (instance, class or property) of the domain ontology.

A paper (respectively a sentence) may be about (respectively may mention) several formal terms. In fact, this last property **isAbout** is defined from the properties **mentions** and **isaSentenceOf** by a rule saying that: ***if*** *a sentence ?s belonging to a paper ?p mentions a formal term*

Manuscript submitted July 2, 2014. This work has been partially supported by the LabEx PERSYVAL-Lab (ANR-11-LABX-0025-01) and by IUF (Institut Universitaire de France).

(1) University Grenoble Alpes, LIG, FR-38000 Grenoble, France (contact e-mail: marie-christine.rousset@imag.fr, others email: firstname.lastname@imag.fr).

(2) University of Crete, Greece.

(3) CHU, INSERM UMR1043 and University of Toulouse III, FR-31062 Toulouse, France (e-mail: tauber.mt@chu-toulouse.fr).

?t ***then*** *it can be inferred that the paper ?p is about the formal term ?t.*
Another rule says that: ***if*** *a sentence ?s mentions a formal term ?t that is either an instance or a subclass of a class ?c* ***then*** *it can inferred that ?s mentions the formal term ?c.*
We have chosen to store and process the resulting knowledge base as a deductive RDF triple-store using, but not limited to, TopBraid Composer [5]. TopBraid Composer is a commercial tool specifically designed for RDF, which is also available as free version. It fully supports the SPARQL query language for expressive querying and also has built-in support for custom inference rules. These rules are applied to infer automatically all the RDF triples that can be entailed from the original triples and the rules.

Step 4 consists in exploiting the resulting knowledge base with reasoning and querying capabilities supported by TopBraid Composer. This knowledge base, made of a set of RDF triples and a set of rules, is a deductive database that can be saturated [3] and then queried using the query language SPARQL [4]. SPARQL is the SQL of RDF: it is a W3C standard allowing to ask possibly complex queries on demand. SPARQL is a powerful structured query language that provides a full set of query operations such as SELECT, JOIN, SORT and aggregation operators based on GROUP-BY. It can be used to compute simple statistics and also to compute answers to queries entered by human experts.

III. Novelty

Compared to existing works on semantic search issued from the Information Retrieval community (such as [7], [8], [9], [10], [11], [12], [13], [14]), the distinguishing point of our approach is to offer a powerful SQL-like query language that enables a unified and powerful approach to express fine-grained queries combining structure, content and semantics. The results returned to the users are not a list of ranked documents but precise fine-grained answers and correlations between formal terms of a specialized domain knowledge and relevant sentences within documents.
The only code that has to be used is generic and is restricted to implement the steps 2 and 3 described above in the methodology. The queries and statistics of interest can be specified declaratively and on demand using SPARQL, and computed by any SPARQL query engine. Exploiting the SPARQL power allowed us to develop a uniform and declarative query-based approach for a fine-grained analysis of a specialized bibliographical corpus.
This approach enables querying capabilities that goes much beyond the concept-based search capabilities offered by semantic search engines such as TextPresso [15] for example, and more generally by information retrieval systems. In addition, in contrast with our approach, the computation of statistics of interest is not supported by (semantic) search engines and must be done by writing Perl or Python scripts or by using dedicated tools.

IV. Ontology-based Description of a Textual Corpus

Our approach is generic but in order to explain it and to show its feasibility, we have focused on a specialized medical domain related to a rare disease (Prader-Willi syndrome [1]) of which one of us is an expert and for which we had access to a corpus of almost 6000 articles representative of the broad spectrum of the literature relevant to this rare syndrome. We first give an overview of the domain ontology and of its automatic enrichment. We then show through examples the power of SPARQL queries both for formulating expert demands of information and for computing some useful statistics on the corpus and its content.

A. Overview of the Prader-Willi ontology

The most general classes are **Anomaly_PW**, **Symptom_PW**, **Troubles_PW**, **Treatment_PW**, **Hormone**, and **Gene** (where the suffix **_PW** denotes that the anomalies, symptoms, etc. that we consider are those specific to Prader-Willi setting). Instances and Subclasses can be declared in RDF by triples built with the RDF property **rdf:type** and the RDFS property **subClassOf**, as shown in Example 1.
Example 1:
The triple *<obesity, rdf:type, Symptom_PW>* is the declaration of the fact that obesity is a symptom encountered in Prader Willi syndrome, while the following triple *<Behaviour_Troubles_PW,rdfs:subClassOf, Troubles_PW* encodes the knowledge that behaviour troubles are a special kind of troubles occurring for patients suffering from Prader Willi syndrome.
Figure 1 shows an extract of the resulting class hierarchy.

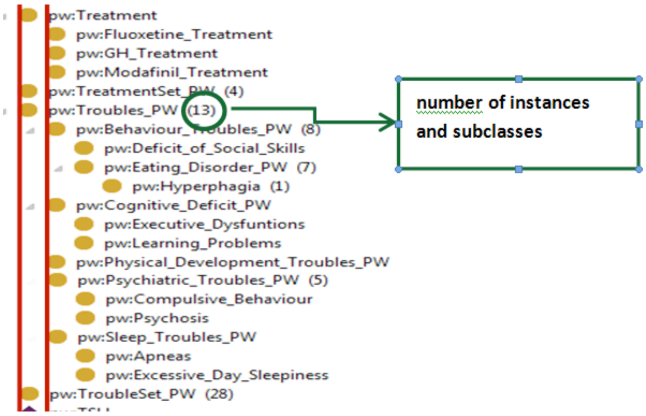


Fig. 1 Hierarchy of classes in the Prader-Willi ontology (extract)

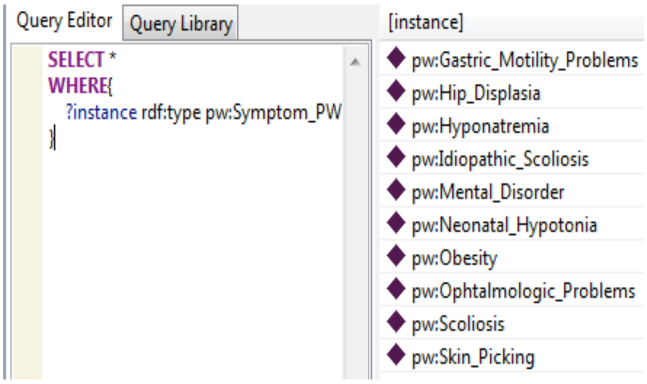
We give below the different domain properties that we have defined:
- **Associated_With:** to relate a symptom to a trouble.
- **Caused_By**: to relate a trouble to an anomaly.
- **Stimulates**: to relate an hormone to another hormone.

Rules enable to express transitivity of the subclass relation and also inheritance of instances between classes like the rule shown in Example 2.

Example 2:
If <?i , rdf:type, ?c1> **and** <?c1, rdfs:subClassOf, ?c2>
Then <?i, rdf:type, ?c2>.
The variables (indicated by ?) are placeholders that can be replaced by constants denoting specific instances and classes accordingly. Such a rule will infer for instance that boulimia (declared as an instance of hyperphagia, itself asserted as a subclass of Eating_Disorder_PW)) is a special case of eating disorder related to Prader Willi syndrome.

Querying the resulting saturated RDF dataset using SPARQL guarantees to be sound and complete, i.e., to return all the answers satisfying any input query given the input RDF facts and the rules.

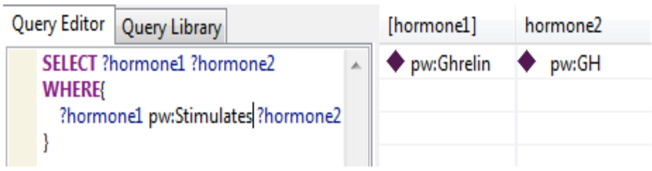
As an illustration, Figures 2 and 3 are screenshots of SPARQL queries and their answers using TopBraid Composer.



```
SELECT *
WHERE{
   ?instance rdf:type pw:Symptom_PW
}
```

[instance]
pw:Gastric_Motility_Problems
pw:Hip_Displasia
pw:Hyponatremia
pw:Idiopathic_Scoliosis
pw:Mental_Disorder
pw:Neonatal_Hypotonia
pw:Obesity
pw:Ophtalmologic_Problems
pw:Scoliosis
pw:Skin_Picking

Fig. 2 SPARQL query asking for symptoms instances in Prader Willi



```
SELECT ?hormone1 ?hormone2
WHERE{
   ?hormone1 pw:Stimulates ?hormone2
}
```

[hormone1]	hormone2
pw:Ghrelin	pw:GH

Fig. 3 SPARQL query asking for pairs of hormones stimulating each other

The current domain ontology is small (25 classes, 61 instances and 3 relations) and far from being complete. It would require more interaction with the expert. Nevertheless, the important point is that it can be easily updated by adding new classes, instances or properties using TopBraid Composer, or by enriching it using medical ontologies available in Linked Data cloud.

B. Incorporating relevant information extracted from the corpus into the domain ontololy

This step relies on a table of correspondence provided by the expert where each formal term of the Prader-Willi ontology is associated with a list of textual expressions that correspond to it in the scientific literature.

Example 3:
The current table of correspondence lists *feeding problems, eating disorders, food intake disorders, appetite regulation problems* as textual expressions corresponding to the formal term *Eating_Disorder_PW* (declared as a class in the domain ontology).
First, for each paper, the sentences that contain one of the textual expressions corresponding to a formal term are automatically extracted.

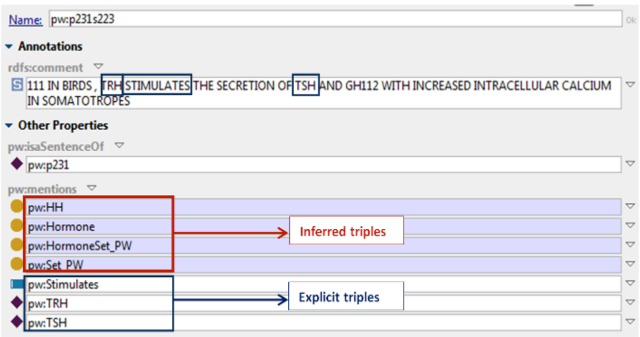
Example 4: From the following inserted triples (where p1608s14 is the identifier of the 14th sentence extracted from the paper number 1608, and HH denotes the class of Hypothalamic Hormones):

<p1608s14, mentions, Oxytocin>,
<Oxytocin, rdf:type, HH> ,
<HH, rdfs:subClassOf, Hormone>

It will be inferred and added to the saturated ontology:

<pw:p1608s14 pw:mentions pw:HH>
<pw:p1608s14 pw:mentions pw:Hormone>

Fig.4 shows a screenshot of the *combined ontology* mapping textual content and formal terms of the domain ontology.



Name: pw:p231s223

Annotations

rdfs:comment
111 IN BIRDS , TRH STIMULATES THE SECRETION OF TSH AND GH112 WITH INCREASED INTRACELLULAR CALCIUM IN SOMATOTROPES

Other Properties

pw:isaSentenceOf
pw:p231

pw:mentions
pw:HH
pw:Hormone
pw:HormoneSet_PW
pw:Set_PW
pw:Stimulates
pw:TRH
pw:TSH

Fig. 4 Screenshot showing the combined ontology

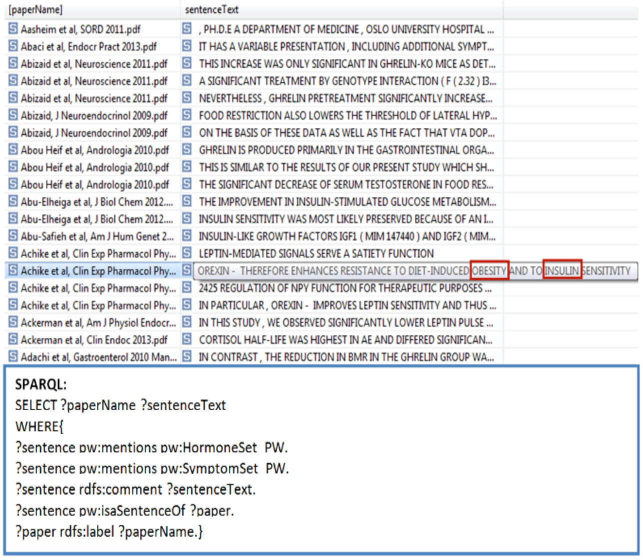
In terms of size, the domain ontology alone is 160 KB while the combined ontology (i.e., the domain ontology enriched with knowledge about the textual content of the corpus) is much bigger: 264 MB corresponding to 1.627.068 RDF triples built on 27 classes, 7 properties and 365.798 instances. The number of scientific papers in the corpus is 5950 from which 359.787 sentences have been extracted (because relevant to at least one formal term of the Prader-Willi ontology). Finally, the size of the saturated combined ontoloy is 393 MB corresponding to 2.941.688 RDF triples.

V. Query-based Analysis and Semantic Search

In this section, we show through examples the power of SPARQL queries both for formulating expert demands and for computing some useful statistics on the corpus and its content. The queries are classified into three categories. The first category illustrates typical queries of experts to search in a corpus information related to their domain of expertise. The second category illustrates how to use queries for a simple statistical analysis of corpus, while third category enables fine-grained semantic analysis of its content.

A. Query-based information retrieval

Fig. 5 shows a query (and its results) asking for papers that contain sentences mentioning both hormones and symptoms, and if such papers exist, to display their titles and the corresponding sentences.
It is worthwhile to note that, thanks to the triples inferred by rules, this general query (in which no specific hormone and no specific symptom is made specific) allows to return very relevant specific information. For instance, the sentence whose full content is displayed in Fig.5 contains **obesity** (that is a specific symptom encountered in Prader-Willi) and **insulin** (that is a specific hormone) that are not explicit in the query. Using search engines based on keyword-based queries, it would have been impossible to express such a query and to obtain such precise answers.

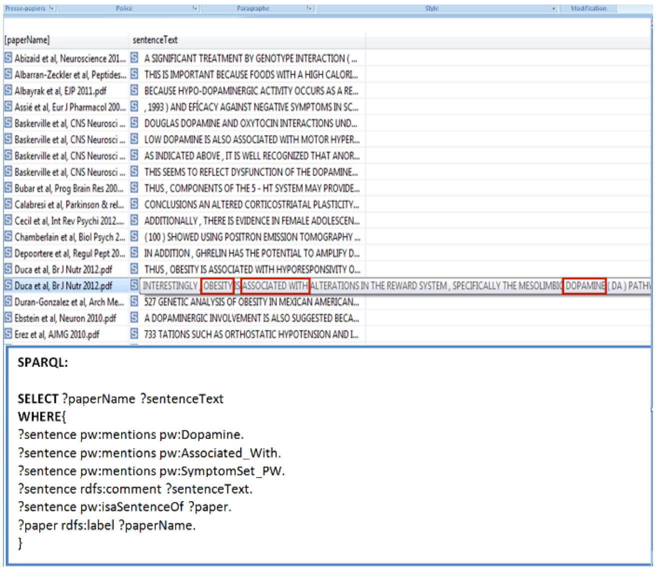


[paperName]	sentenceText
Aasheim et al, SORD 2011.pdf	, PH.D.E A DEPARTMENT OF MEDICINE , OSLO UNIVERSITY HOSPITAL ...
Abaci et al, Endocr Pract 2013.pdf	IT HAS A VARIABLE PRESENTATION , INCLUDING ADDITIONAL SYMPT...
Abizaid et al, Neuroscience 2011.pdf	THIS INCREASE WAS ONLY SIGNIFICANT IN GHRELIN-KO MICE AS DET...
Abizaid et al, Neuroscience 2011.pdf	A SIGNIFICANT TREATMENT BY GENOTYPE INTERACTION (F (2.32) B...
Abizaid et al, Neuroscience 2011.pdf	NEVERTHELESS , GHRELIN PRETREATMENT SIGNIFICANTLY INCREASE...
Abizaid, J Neuroendocrinol 2009.pdf	FOOD RESTRICTION ALSO LOWERS THE THRESHOLD OF LATERAL HYP...
Abizaid, J Neuroendocrinol 2009.pdf	ON THE BASIS OF THESE DATA AS WELL AS THE FACT THAT VTA DOP...
Abou Heif et al, Andrologia 2010.pdf	GHRELIN IS PRODUCED PRIMARILY IN THE GASTROINTESTINAL ORGA...
Abou Heif et al, Andrologia 2010.pdf	THIS IS SIMILAR TO THE RESULTS OF OUR PRESENT STUDY WHICH SH...
Abou Heif et al, Andrologia 2010.pdf	THE SIGNIFICANT DECREASE OF SERUM TESTOSTERONE IN FOOD RES...
Abu-Elheiga et al, J Biol Chem 2012....	THE IMPROVEMENT IN INSULIN-STIMULATED GLUCOSE METABOLISM...
Abu-Elheiga et al, J Biol Chem 2012....	INSULIN SENSITIVITY WAS MOST LIKELY PRESERVED BECAUSE OF AN I...
Abu-Safieh et al, Am J Hum Genet 2...	INSULIN-LIKE GROWTH FACTORS IGF1 (MIM 147440) AND IGF2 (MIM...
Achike et al, Clin Exp Pharmacol Phy...	LEPTIN-MEDIATED SIGNALS SERVE A SATIETY FUNCTION
Achike et al, Clin Exp Pharmacol Phy...	OREXIN - THEREFORE ENHANCES RESISTANCE TO DIET-INDUCED OBESITY AND TO INSULIN SENSITIVITY
Achike et al, Clin Exp Pharmacol Phy...	2425 REGULATION OF NPY FUNCTION FOR THERAPEUTIC PURPOSES ...
Achike et al, Clin Exp Pharmacol Phy...	IN PARTICULAR , OREXIN - IMPROVES LEPTIN SENSITIVITY AND THUS ...
Ackerman et al, Am J Physiol Endocr...	IN THIS STUDY , WE OBSERVED SIGNIFICANTLY LOWER LEPTIN PULSE ...
Ackerman et al, Clin Endoc 2013.pdf	CORTISOL HALF-LIFE WAS HIGHEST IN AE AND DIFFERED SIGNIFICAN...
Adachi et al, Gastroenterol 2010 Man...	IN CONTRAST , THE REDUCTION IN BMR IN THE GHRELIN GROUP WA...

```
SPARQL:
SELECT ?paperName  ?sentenceText
WHERE{
?sentence pw:mentions pw:HormoneSet_PW.
?sentence pw:mentions pw:SymptomSet_PW.
?sentence rdfs:comment ?sentenceText.
?sentence pw:isaSentenceOf ?paper.
?paper rdfs:label ?paperName.}
```

Fig. 5 How to find papers with sentences mentioning some hormones and some symptoms related to Prader-Willi

Fig.6 shows a query that is more specific than the previous one in which the expert is interested in finding papers in which a same sentence mentions explicitly an association between a symptom and a specific **neurotransmitter** (namely the dopamine).



[paperName]	sentenceText
Abizaid et al, Neuroscience 201...	A SIGNIFICANT TREATMENT BY GENOTYPE INTERACTION (...
Albarran-Zeckler et al, Peptides...	THIS IS IMPORTANT BECAUSE FOODS WITH A HIGH CALORI...
Albayrak et al, EJP 2011.pdf	BECAUSE HYPO-DOPAMINERGIC ACTIVITY OCCURS AS A RE...
Assié et al, Eur J Pharmacol 200...	, 1993) AND EFÍCACY AGAINST NEGATIVE SYMPTOMS IN SC...
Baskerville et al, CNS Neurosci ...	DOUGLAS DOPAMINE AND OXYTOCIN INTERACTIONS UND...
Baskerville et al, CNS Neurosci ...	LOW DOPAMINE IS ALSO ASSOCIATED WITH MOTOR HYPER...
Baskerville et al, CNS Neurosci ...	AS INDICATED ABOVE , IT IS WELL RECOGNIZED THAT ANOR...
Baskerville et al, CNS Neurosci ...	THIS SEEMS TO REFLECT DYSFUNCTION OF THE DOPAMINE...
Bubar et al, Prog Brain Res 200...	THUS , COMPONENTS OF THE 5 - HT SYSTEM MAY PROVIDE...
Calabresi et al, Parkinson & rel...	CONCLUSIONS AN ALTERED CORTICOSTRIATAL PLASTICITY...
Cecil et al, Int Rev Psychi 2012....	ADDITIONALLY , THERE IS EVIDENCE IN FEMALE ADOLESCEN...
Chamberlain et al, Biol Psych 2...	(100) SHOWED USING POSITRON EMISSION TOMOGRAPHY ...
Depoortere et al, Regul Pept 20...	IN ADDITION , GHRELIN HAS THE POTENTIAL TO AMPLIFY D...
Duca et al, Br J Nutr 2012.pdf	THUS , OBESITY IS ASSOCIATED WITH HYPORESPONSIVITY O...
Duca et al, Br J Nutr 2012.pdf	INTERESTINGLY , OBESITY IS ASSOCIATED WITH ALTERATIONS IN THE REWARD SYSTEM , SPECIFICALLY THE MESOLIMBIC DOPAMINE (DA) PATH
Duran-Gonzalez et al, Arch Me...	527 GENETIC ANALYSIS OF OBESITY IN MEXICAN AMERICAN...
Ebstein et al, Neuron 2010.pdf	A DOPAMINERGIC INVOLVEMENT IS ALSO SUGGESTED BECA...
Erez et al, AJMG 2010.pdf	733 TATIONS SUCH AS ORTHOSTATIC HYPOTENSION AND I...

```
SPARQL:

SELECT ?paperName ?sentenceText
WHERE{
?sentence pw:mentions pw:Dopamine.
?sentence pw:mentions pw:Associated_With.
?sentence pw:mentions pw:SymptomSet_PW.
?sentence rdfs:comment ?sentenceText.
?sentence pw:isaSentenceOf ?paper.
?paper rdfs:label ?paperName.
}
```

Fig. 6 How to find papers with sentences mentioning an association between a Prader-Willi symptom and dopamine.

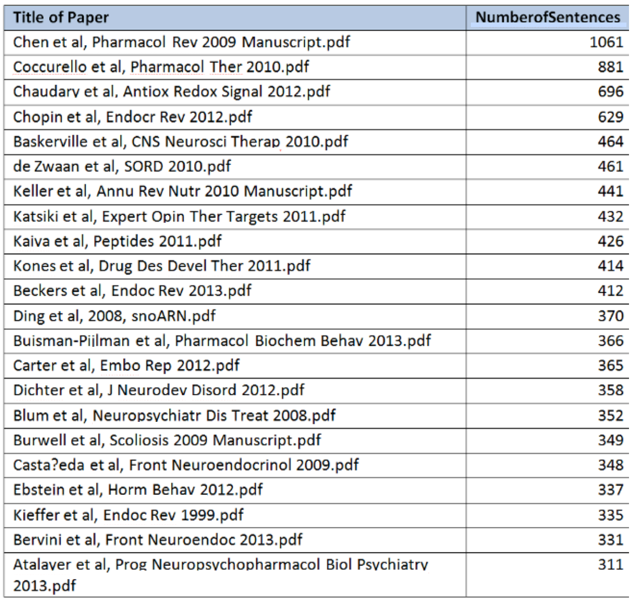
These two queries clearly illustrate the added-value of our ontology-based approach equipped with querying and inference capabilities compared to keyword-based or even concept-based search.

B. *Query-based simple statistical analysis of the corpus*

First, we have measured to what extent the corpus is related to the Prader-Willi ontology described in Section IV.

Fig.7 and Fig.8 show that, despite the small size of the current Prader-Willi domain ontology, our approach allowed to automatically extract from the corpus a significant amount of information relevant to the Prader-Willi domain. Fig.7 shows the result of the following SPARQL query that asks for displaying the 30 papers that contain the most sentences related to the current Prader-Willi ontology.

```
Sparql:
SELECT ?titlePaper (COUNT(*) AS ?count)
WHERE{
   ?a pw:isaSentenceOf ?s.
   ?s rdfs:label ?titlePaper.
}
GROUP BY ?s ?titlePaper
ORDER BY DESC(?count)
LIMIT 30
```



Title of Paper	NumberofSentences
Chen et al, Pharmacol Rev 2009 Manuscript.pdf	1061
Coccurello et al, Pharmacol Ther 2010.pdf	881
Chaudary et al, Antiox Redox Signal 2012.pdf	696
Chopin et al, Endocr Rev 2012.pdf	629
Baskerville et al, CNS Neurosci Therap 2010.pdf	464
de Zwaan et al, SORD 2010.pdf	461
Keller et al, Annu Rev Nutr 2010 Manuscript.pdf	441
Katsiki et al, Expert Opin Ther Targets 2011.pdf	432
Kaiva et al, Peptides 2011.pdf	426
Kones et al, Drug Des Devel Ther 2011.pdf	414
Beckers et al, Endoc Rev 2013.pdf	412
Ding et al, 2008, snoARN.pdf	370
Buisman-Pijlman et al, Pharmacol Biochem Behav 2013.pdf	366
Carter et al, Embo Rep 2012.pdf	365
Dichter et al, J Neurodev Disord 2012.pdf	358
Blum et al, Neuropsychiatr Dis Treat 2008.pdf	352
Burwell et al, Scoliosis 2009 Manuscript.pdf	349
Casta?eda et al, Front Neuroendocrinol 2009.pdf	348
Ebstein et al, Horm Behav 2012.pdf	337
Kieffer et al, Endoc Rev 1999.pdf	335
Bervini et al, Front Neuroendoc 2013.pdf	331
Atalayer et al, Prog Neuropsychopharmacol Biol Psychiatry 2013.pdf	311

Fig. 7 Titles of papers found with the most sentences related to Prader-Willi

Fig.8 shows the distribution of the number of the relevant sentences among the papers of the corpus. Around 98% of the papers corpus contain from 1 to 100 sentences mentioning at least one of the formal terms of our ontology, and 1.5% of papers contain between 101 and 200 such sentences. The Prader-Willi entities with the corresponding expressions can be matched to more than 100 sentences in average. This is a very encouraging results considering the limitation of our domain ontology so far. It can be assumed that with a more complete ontology the number of sentences found relevant to the ontology will increase significantly.

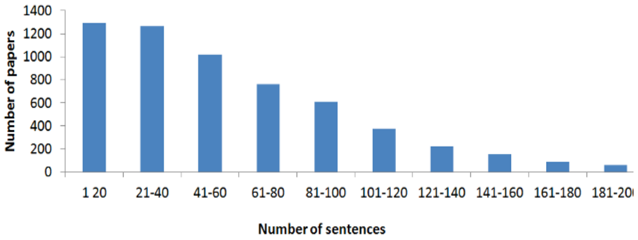
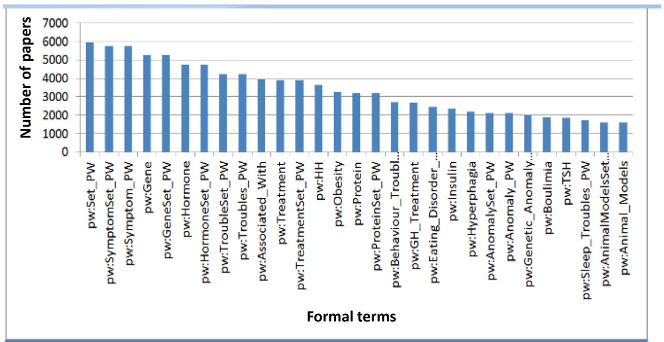


Fig. 8 Histogram showing the distribution of the number of sentences within papers

Second, we have identified the most frequent formal terms that are mentioned in the papers. The query and the corresponding results are given in Fig. 9.



```
SPARQL:
SELECT ?formalTerm ?freq_formal_term
WHERE{
SELECT (COUNT(*) AS ?freq_formal_term) ?formalTerm
WHERE
{ ?paper pw:isAbout ?formalTerm.
}
GROUP BY ?formalTerm }
ORDER BY DESC(?freq_formal_term)
LIMIT 30
```

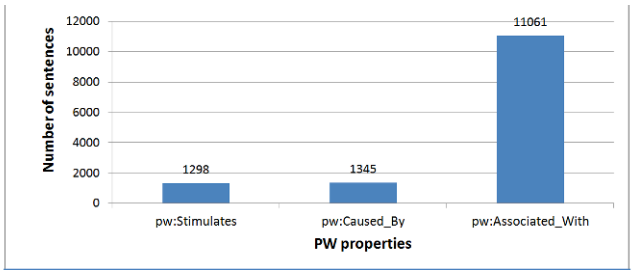
Fig. 9 The query asking for the 30 most frequent formal terms mentioned the papers, and the visualization of its results.

Not surprisingly, the most frequent formal terms that are mentioned in the papers are the general entities (Symptom, Gene, Hormone, Troubles). This is because our approach is able to infer that a paper implicitly mentions a given formal term (e.g., a Symptom) if it mentions explicitly a subclass of an instance of it. Interestingly, among the specific formal terms (i.e., instances in the ontology), **Obesity** is the most frequent in the papers of the Prader-Willi corpus. This is a strong hint (validated by the expert) that obesity is one of the main symptoms of Prader-Willi syndrome.

C. Query-based fine-grained analysis of the corpus

We have written several queries to investigate co-occurrences within a same sentence of different types of formal terms related to the Prader-Willi domain.

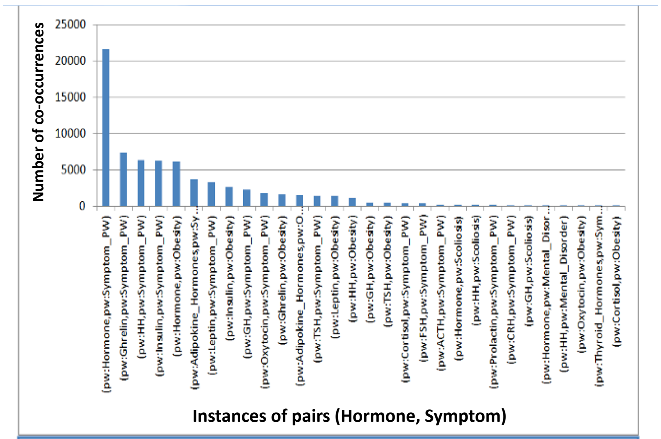
Fig. 10 reports on the co-occurrences that have been queried and found between one of the 3 domain specific relations and the one hand, and a formal concept of the domain on the other hand. The most frequent relation co-occurring with other concepts related to Prader-Willi is clearly the relation **Associated_With**. Exploring the corresponding sentences could help to find new concepts to enrich the current ontology.



```
SELECT (COUNT(*) AS ?numSentences) ?property
WHERE{{
?property rdf:type owl:ObjectProperty.
?s pw:mentions ?property.}
?s pw:mentions pw:Set_PW.}
GROUP BY ?property
```

Fig. 10 The query-based computation of the number of co-occurrences of each of the 3 relations with a formal concept

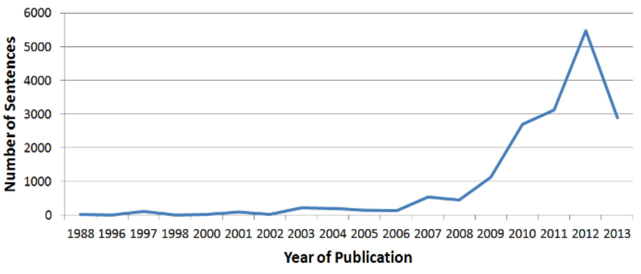
Fig.11 shows the query for computing the 30 most frequent co-occurrences of symptoms and hormones, and the results ordered by their frequency. As we can see, the number of (explicit or implicit) co-occurrences of symptoms and hormones is big (approximately 21.000 sentences), which suggests a strong relation between them. The most frequent co-occurrence between two specific entities respectively instances of the classes Hormones and Symptoms is **Insulin** with **Obesity**. This suggests a possible link between insulin and obesity in Prader-Willi domain.



```
SPARQL:
SELECT (COUNT(*) AS ?numberOfOccurence)?hormone ?symptom
WHERE{
?hormone rdf:type pw:HormoneSet_PW.
?symptom rdf:type pw:SymptomSet_PW.
?s pw:mentions ?hormone.
?s pw:mentions ?symptom.
FILTER(?hormone!=?symptom)
}
GROUP BY ?hormone ?symptom
ORDER BY DESC(?numberOfOccurence)
LIMIT 30
```

Fig. 11 The query-based computation of the number of co-occurrences of hormones and symptomes

Among these correlation results, the expert wanted to zoom on the correlation between a specific hormone (namely **Oxytocin**) and the general formal term **Symptom_PW** denoting the set of symptoms of the Prader-Willi declared in our ontology. Her demand was then to visualize the evolution over time of the correlation of oxytocin with Prader-Willi syndrome. Fig.12 shows the query corresponding to this demand, and the resulting curve.



```
SPARQL:
SELECT (COUNT(?paper) AS ?numOfPapers) ?yearOfPublication
WHERE{
?sentence pw:mentions pw:Oxytocin.
?sentence pw:isaSentenceOf ?paper.
?paper pw:publishedIn ?yearOfPublication.
}
GROUP BY ?yearOfPublication
ORDER BY ASC(?yearOfPublication)
```

Fig. 12 Query-based computation of evolution over time of occurrences of oxytocyn into the corpus

Fig. 12 clearly shows the increasing importance taken by oxytocyn within the scientific literature about Prader-Willi syndrome since 2006. This reflects the current advances on this rare disease for which the role of oxytocyn has been investigated quite recently. Note that the decreasing of the curve for 2013 may be due to the fact that our corpus contains much less papers published in 2013 than in the previous years.

VI. Conclusion

We have designed a novel approach to filter and query a corpus of scientific articles related to a domain of specialization for which an ontology exists or can be built. The novelty of this approach is to represent the textual content relevant to the domain of interest as knowledge added to the ontology, thus making explicit the link between sentences and papers with formal terms they refer to. This enables to use a powerful SQL-like query language, namely SPARQL to express possibly complex queries on demand. Our approach is declarative and generic. The domain ontology can be simply updated by an expert who just has to add or delete triples using the editor of TopBraid Composer, and additional articles can be added to the corpus but the whole algorithmic machinery remains unchanged. In this paper, this approach is illustrated on a specialized medical domain, namely the Prader-Willi syndrome. However, it can be easily applied to another domain simply by providing as input an ontology of this domain and a corpus of articles.

We have implemented this approach using W3C standards that make it conform to Linked Data [6]. This opens the possibility to easily extend it to ontologies available in the Linked Data cloud.

As future work, in line with existing works combining ontologies and text mining (e.g., [15], [16], [18], [19]), we plan to exploit text mining techniques in order to (semi)-automatically enrich the domain ontology. The fact to have a preliminary ontology is important because it can be exploited to facilitate concept extraction. For instance, by focusing on sentences recognized by our method as mentioning one relation (e.g, Associated_With), extracting expressions corresponding to new concepts based on their position in the sentence compared to the expressions known to refer to the recognized relation is a simple but promising approach. In turn, the enriched ontology will enable to extract more sentences relevant to the domain of interest. This iterative process, combining simple text mining techniques and ontology-based corpus filtering is likely to be robust while limiting the manual help of human experts.

References

[1] Cassidy,Driscoll. Prader-Willi syndrome. European Journal of Human Genetics (2009) 17:3–13

[2] Hayes P: RDF Semantics. Recommendation, World Wide WebConsortium 2004. [http://www.w3.org/TR/2004/REC-rdf-mt-20040210

[3] Abiteboul S. , Manolescu I., Rigaux P., Rousset M-C, Senellart P. Web Data Management, Cambridge University Press,2012.

[4] Prud'hommeaux E, Seaborne A: SPARQL query language for RDF.Latest version available as http://www.w3.org/TR/rdf-sparql-query/ 2008. [http://www.w3.org/TR/2008/REC-rdf-sparql-query-20080115

[5] TopBraid Composer [http://www.topquadrant.com/]

[6] Linked Data [http://linkeddata.org/]

[7] Malo, Pekka, et al. "Semantic content filtering with Wikipedia and ontologies." Data Mining Workshops (ICDMW), 2010 IEEE International Conference on. IEEE, 2010.

[8] Perez-Iratxeta, Carolina, et al. "Update on XplorMed: a web server for exploring scientific literature." Nucleic Acids Research 31. 13 (2003): 3866-3868.

[9] Webber, William. "Evaluating the Effectiveness of Keyword Search." IEEE Data Eng. Bull. 33.1 (2010): 54-59.

[10] Qiu, Yonggang, and Hans-Peter Frei. "Concept based query expansion." Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 1993.

[11] Santos, José Carlos Almeida, and Manuel Fonseca de Sam Bento Ribeiro. "Improving search engine Query Expansion techniques with ILP."

[12] Perez-Iratxeta, Carolina, et al. "Update on XplorMed: a web server for exploring scientific literature." Nucleic Acids Research 31.13 (2003): 3866-3868.

[13] Tsuruoka, Yoshimasa, Jun'ichi Tsujii, and Sophia Ananiadou. "FACTA: a text search engine for finding associated biomedical concepts." Bioinformatics 24.21 (2008): 2559-2560.

[14] Egozi, Ofer, Evgeniy Gabrilovich, and Shaul Markovitch. "Concept-Based Feature Generation and Selection for Information Retrieval." AAAI. 2008.

[15] Müller, Hans-Michael, Eimear E. Kenny, and Paul W. Sternberg. "Textpresso: an ontology-based information retrieval and extraction system for biological literature." PLoS biology 2.11 (2004)

[16] Spasic, Irena, et al. "Text mining and ontologies in biomedicine: making sense of raw text." Briefings in bioinformatics 6.3 (2005): 239-251.

[17] Gregorowicz, Andrew, and Mark A. Kramer. "Mining a large-scale term-concept network from wikipedia." MITRE Corporation 202 (2006).

[18] Maier, Holger, et al. "LitMiner and WikiGene: identifying problem-related key players of gene regulation using publication abstracts." Nucleic acids research 33.suppl 2 (2005): W779-W782.

[19] Cheng, Dean, et al. "PolySearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites." Nucleic acids research 36.suppl 2 (2008): W399-W405.