Master of Computer Science

University of Crete and
University of Joseph Fourier

Semantic Filtering of Bibliographical Articles

Elena Michael

Research Project performed at
LIG (Laboratoire d'Informatique de Grenoble)
HADAS Team (Heterogenous Autonomous distributed DAta Services)

Under the supervision of
Marie-Christine Rousset
Co-supervisors: Prof. Cyril Labbé, Prof. Fabrice Jouanot and Dr.Federico Ulliana

Defended before a jury composed of:
Prof. Catherine Berrut
Dr. Noha Ibrahim
Prof. Jerome Euzenat/ Prof. Eric Gaussier
Prof. Eric Gaussier
Dr. Jean-Marc Vincent
Dr. Jerome David

<June > <2014>

University of Joseph Fourier
Computer Science Department
Semantic filtering of Bibliographical Articles

Thesis submitted by
Elena Michael
in partial fulfillment of the requirements for the
Masters' of Science degree in Computer Science of University Joseph Fourier
University of Crete
THESIS APPROVAL

Author:

Elena Michael

Committee approvals:

Marie-Christine Rousset
(Professor, Thesis Supervisor)

Prof. Catherine Berrut

Dr. Noha Ibrahim

Prof. Jerome Euzenat/ Prof. Eric Gaussier

Dr. Jean-Marc Vincent

Dr. Jerome David
(External Reviewer)

Presented on 23 of June 2014, Grenoble

Abstract

For researchers, finding bibliographical articles helps them to improve their knowledge on their domain of expertise. Hence, it's a crucial but time-consuming task. In this project, we have investigated 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 ontology. More precisely, we have developed an ontology related to a rare disease (the Prader-Willi syndrome) as a deductive database that we have described using Semantic Web standards (namely, RDFS enriched with rules), and stored as a set of RDF triples. We have obtained a knowledge base describing the content of a corpus filtered by the terms of the domain ontology. This knowledge base, made of a set of RDF triples and a set of rules, can be seen as a deductive database that can be saturated and then queried using the query language SPARQL. In particular, this approach enables querying capabilities that goes much beyond the keyword-based search capabilities offered by search engines and more generally by information retrieval systems. We have shown by example the interest and the power of such a declarative knowledge-based approach both for computing variety of statistics on the corpus and its correlation with formal terms of a domain of interest, and for helping experts to find useful fine-grained information within a textual corpus.

Acknowledgments

I take this opportunity to express my profound gratitude to my supervisor Mrs Marie-Christine Rousset (Prof. of Joseph Fourier, member of Laboratory Informatique Grenoble (LIG)) who gave me the chance to come in the HADAS (Heterogenous Autonomous distributed DAta Services) team of LIG. I would like to express my deep regards and appreciation for her support, exemplary guidance, monitoring and valuable knowledge, which helped me in completing this thesis.

I would especially like to thank Prof. Cyril Labbé, Prof. Fabrice Jouanot and Doc.Federico Ulliana for their kind co-operation, coaching and guidance during my master thesis. This collaboration gave me opportunity to understand more the process of research and to improve my skills.

I would like to convey my gratefulness to my supervisor and the whole team for this experience. The skills, the knowledge and advices which I have gained throughout my thesis I perceive as very valuable component in my future career development.

Moreover, I place a deep sense of gratitude to my parents, Keti and Alexandros, for their valuable support to me all these years of my studies. Also my brother, Stavros, who kept me interested in the computer science field as he also studies the same field as me. Especially, I want to thank my spiritual priest who gave me courage to face my difficulties.

Last but not least, I want to express a deep sense of gratitude to my friends from Greece, Rodoula, Nelli, Lina, Myron, Kostantinos and Argyri for their constant encouragement to continue my studies. I want to thank a lot Sophia and Andreas for helping me to adapt here. Also, it was a very great experience for me here in Grenoble because I made new friends and they helped me to spend my master period here enjoyfully. I want to thank for their help: Rishabh, Meriam, Hicham, Anil and also my friends from LIG: Behrooz, Shadi, Sahar, Thiago, Shaswat, Bilyana, Oleg, and Uzma.

CONTENTS

FIGURES

1. 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 project, we have investigated 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.

More precisely, we have developed an ontology related to a rare disease (the Prader-Willi syndrome) as a deductive database that we have described using Semantic Web standards (namely, RDFS enriched with rules), and stored as a set of RDF triples. We have used this ontology to extract from a corpus of articles related to this rare disease, for each article, the set of sentences mentioning at least one formal term of the ontology. We have then enriched the set of RDF triples by new triples relating each identifier of the extracted sentences to the formal term(s) they mention, and each paper identifier to the identifiers of sentences that it contains. As a result, we have obtained a knowledge base describing the content of a corpus filtered by the terms of a domain ontology. This knowledge base, made of a set of RDF triples and a set of rules, can be seen as a deductive database that can be saturated and then queried using the query language SPARQL. SPARQL is the SQL of RDF is a W3C standard allowing to ask possibly complex queries and to obtain precise and answers.

We have shown by example the interest and the power of such a declarative knowledge-based approach both for computing varied statistics on the corpus and its correlation with formal terms of a domain of interest, and for helping experts to find useful fine-grained information within a textual corpus.

1.1. Research Setting and Problem Statement

In medicine, and in particular for rare diseases that are ill-known, scientific advances often come from analogies with the physiopathology of other diseases having some common symptoms, and for which the causes or the involved genes and mechanisms are better known. This requires to continuously keeping up to date with the latest results in the literature that are related to the in-depth knowledge on the disease of interest, directly or indirectly through biological mechanisms likely to be complex and only known by domain experts.

The setting of this project is the Prader-Willi syndrome [4], a rare disease for which we have collaborated with an expert (Pr. Maité Tauber, Centre de physiopathologie de Toulouse-Purpan CHU de Toulouse) who provided also a corpus of bibliographical articles around 6000 papers that are representative of the broad spectrum of the literature relevant for a better understanding of this rare syndrome.

The problem that we address in this project is how to improve the accuracy of retrieving specialized information within a textual scientific corpus, compared to standard keyword-based information retrieval.

1.2. Summary of the Proposed Approach

Our approach consists in using a domain ontology as a pivot structured and specialized vocabulary between users (expert in this domain) and a bibliographical corpus. First, the ontology is used as a filter for automatically extracting the sentences in the papers of the corpus that mention formal terms of the ontology. Second, the domain ontology and the filtered content of the corpus are stored together within a knowledge base equipped with reasoning and querying capabilities. We have chosen to store and process the resulting knowledge base as a deductive RDF triple-store using TopBraid Composer. 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.

Querying using SPARQL the resulting saturated RDF dataset 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. Exploiting the SPARQL power allowed us to develop a uniform and declarative query-based approach for a fine-grained analysis and semantic search of a specialized bibliographical corpus. In particular, this approach enables querying capabilities that goes much beyond the keyword-based search capabilities offered by search engines and more generally by information retrieval systems.

For instance, when the user submits keywords into Google search engine, it returns a set of results ranked by the similarity metrics. Afterwards, user needs to navigate these links in order to check if the content is relevant according to his interest. This is time-consuming and undesirable for the user.

The following picture illustrates summarized the proposed approach that we have implemented and evaluated?

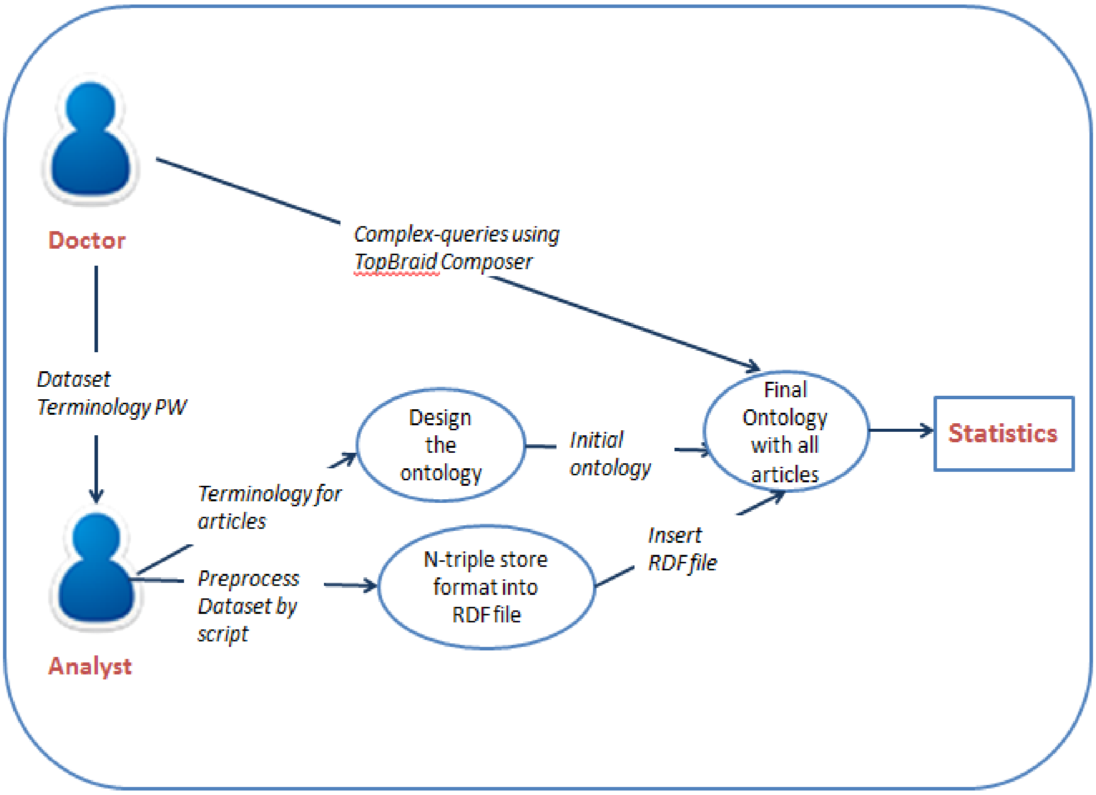

Doctor
Complex-queries using TopBraid Composer
Dataset Terminology PW
Design the ontology
Initial ontology
Final Ontology with all articles
Statistics
Terminology for articles
Insert RDF file
Preprocess Dataset by script
N-triple store format into RDF file
Analyst

Figure 1.1:Representation of our approach

1.3. Contributions of this research project

We have made three main contributions that are described in the three following sections. They can be summarized as follows.

1. *Design of a Prader-Willi ontology*
   With the help of Prof. M. Tauber, we have identified the main medical and biological concepts involved in the Prader-Willi syndrome, the (structural or causal) relations existing between them, and also some of their instances. We have modeled this ontology as a RDF deductive dataset made of RDF triples and rules. We have used TobBraid Composer to edit it, store it, apply the rules to compute the saturated RDF dataset, and to query it using SPARQL. At the moment, the current preliminary version of the ontology contains 27 classes, 61 instances, 5 relations and 11 rules. This version is far from being complete but it can easily be enriched with new classes, new instances or new properties simply by adding RDF triples or rules.

2. *Automatic filtering and extraction of sentences related to the ontology*
   We have designed and implemented in Perl an automatic processing line for extracting from a corpus of pdf files its textual content that is related to the ontology given as input. Each pdf file is converted into text, from which we retain only the sentences that contain expressions corresponding to formal terms of the ontology. These expressions are predefined by the expert for whom it is quite easy to provide most of the possible alternative ways encountered in scientific articles to express each term of the Prader-Willi ontology she also entered.

3. *Query-based analysis of the corpus.*
   We have modeled the corpus (its structure and its content) as a set of facts and rules relating identifiers of papers and identifiers of sentences through the relation “contains” (relating a paper identifier to the identifiers of the sentences it contains), the relation “mentions” (relating each sentence to the formal terms of the ontology it mentions), and the inferred relation “isAbout” (relating each paper to the formal terms of the ontology it is about). All these facts and rules are added to the RDF deductive dataset, thus resulting in a knowledge base capturing in a uniform way a domain specific knowledge and the structure and content of a corpus of documents related to this domain. Then, we exploit inference and querying capabilities of any semantic web tool (such as TopBraid Composer) to specify and compute statistics on the corpus on demand of the expert.

2. Related Work

Standard information retrieval (IR) techniques are based on representing texts as bag-of-words (BOW) and on keyword-based queries. The general approach is to retrieve the documents that contain (part of) the keywords of the query and to rank them based on similarity metrics. In this section, we will focus on the recent existing works in IR that extend bag of words with semantic annotations. Semantic annotation is a specific metadata generation which helps to provide new information by usage schema [19].

These works are based on semantic annotation to accomplish indexing and retrieval using both traditional IR queries and ontology-based queries. Most of the approaches are focused mostly on implicit query expansion. Besides, there are some approaches that proceed to explicit query expansion by formulating the query from user's textual statements or by defining query patterns.

Query expansion is the process to reformulate the initial format of the query in order to enhance the information retrieval operations. It is accomplished by adding new terms into initial query which are helpful to increase the quality of the results, by virtue of matching additional related documents. In a nutshell, the returning result after the query expansion confirms high retrieval effectiveness regarding with similarity metrics (recall, precision and usefulness). Consequently, the results from the query expansion increase the relevance contrary to initial query. Query expansion occurs either by explicit or by implicit manner.

Firstly, explicit query expansion means when you add new terms that have the same meaning with the initial keywords of the query. Some methods to acquire these new terms are: finding synonyms, finding all the morphological variations by stemming each word, proposing correct terms in case the original terms have spelling errors. In order to select carefully additional search terms for automatic query expansion, it should rely on statistical co-occurrence data.

Recently, a new IR methodology makes use of Explicit Semantic Analysis (ESA). It builds a feature generator using Wikipedia articles as concepts and makes text categorization and improves the BOW text representation [11]. The novelty of this method is the concept encoding. It matches the textual content with the entire collection of available concepts in order to make semantic annotation regarding with the level of relation. Apart from this, there are some other alternative methods: use of document classification, use of syntactic context or use of relevant information [8].

All these approaches combine both explicit and implicit query expansion by using semantic or syntactic analysis correspondingly. Implicit query expansion is the process to add concepts to the query, that are strongly related with the initial query terms. These concepts have actual semantic meaning with the initial terms of the query. Hence, implicit expansion gives benefits to add documents into query's output that are not primarily attached, thereby gaining more precise results.

Some of existing works provide similar functionality as text search engines using semantic annotation in order to rank the output. Users can type keywords queries using either simple free keywords, either Boolean combinations of keywords or concepts. For instance, FACTA is a text search engine which is responsible for assisting users to browse biomedical concepts and retrieve documents, particularly in MEDLINE abstracts by queries. FACTA annotates the documents using predefined concepts that are collected from different resources(biomedical databases and thesauri such as BioThesaurus),and are recognized by dictionary matching [12]. Similar functionality is observed in: LitMiner, WikiGene [14] and XplorMed [15].

Another approach is to allow words from the document to be queried with the purpose to formulate semantic queries. For example, Textpresso[13] is a text processing system which splits papers into sentences. Afterwards, the system splits sentences into words or phrases in order to label them using eXtensible Markup Language (XML) according to the lexicon of ontology. In addition to terms from Textpresso, the ontology contains also the terms from Gene Ontology (The Gene Ontology Consortium 2000) which includes biological terms and synonyms.

Textpresso indexes all sentences regarding with desired labels that contain keywords with boolean operators (AND/OR) to improve retrieval for sentences. The user interface of Textpresso includes a query interface, which provides two different kinds of retrieval, simple and advanced. The user can specify the frequency of terms' occurrences and the text categorization specifying which part of the article is to be searched (title, abstract, body).

There is also another approach which is to formulate queries depending on user's textual statement like in the Wiki-SR framework [5].Wiki-SR framework is an interactive application where the user can submit his interest by giving a "topic statement". This "topic statement" consists of a textual description to define the concerned concepts and also a set of documents that are relevant or irrelevant with the given topic. So, Wiki-SR framework deploys this user's information through Wikipedia's concept-relatedness information. Afterwards, it builds semantic rules or otherwise called "semantified" boolean queries. These Wikipedia-based semantic rules (Wiki-SR rule) represent the user's information preference as an expression of an ordinary boolean query with only difference that the keywords are the terms of Wikipedia and ontology concepts. Also, they serve the purpose to make document filtering. They evaluate whether the query's concepts are mentioned in the document by virtue of Wikipedia concept-related information. They similarly construct semantic measures between any pair of Wikipedia's concepts to ensure effective document filtering and classification performance.

Another approach to express queries is using query patterns for the queries. For example, PolySearch[16], biomedical text mining and web-based tool, text search engines don't offer only Boolean representation to express a query. For instance, PolySearch [16] is a biomedical text mining and web-based tool where the user can define such queries. The specific structure of almost every query in PolySearch is "Given X, find all associated Y's'", where X or Y would be any predefined biomedical classes (human disease, gene, protein etc.). User has ability to apply the query into multiple databases (PubMed, DrugBank, Swissport etc.) and after to rank the returning information.

There are some limitations on existing works. First of all, the user has limits to express his queries because he can express either by Boolean combinations using concepts and simple keywords or by specific query format as PolySearch web server [16]. Moreover, in order to avoid time-consumption, some approaches focus the search exclusively only in abstracts. For example, search and retrieval of LitMiner software [14] is focused only on Pubmed abstracts and the same method is followed also by XplorMed web tool [15]. Finally, the query's results are needed to apply to some calculations using similarity metrics in order to rank them.

To complement existing works, we have designed an efficient approach which is responsible for extracting specialized information from bibliographical articles. Subsequently, we incorporate this information into the domain ontology. Using TBC software, user has ability to define complex queries via SPARQL editor instead of expressing keyword queries.

SPARQL is a powerful structured query language because it provides a full set of query operations such as JOIN, SORT etc. Thus, user can access more easily the Knowledge Base. Moreover, our approach relies not only on the abstracts from scientific papers ((FACTA[12], Update on XplorMed[15])), but on the whole content of the papers. Finally, the output of SPARQL queries is a set of precise answers and it doesn't need to apply information retrieval operations, ranking and indexing, just as the previous approaches. As a result, there is not any time-consumption to process the results from query.

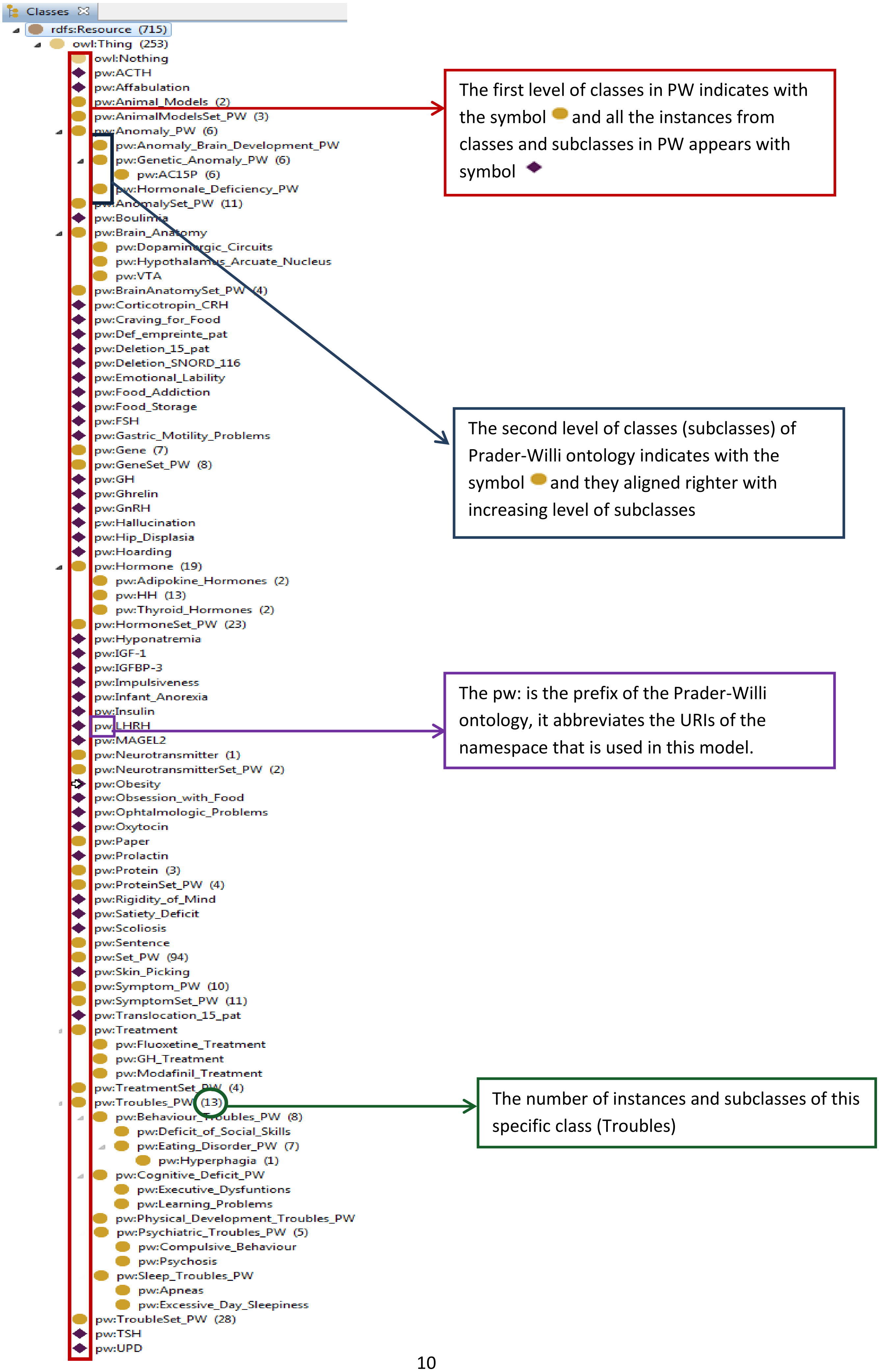
3. Design of a Prader-Willi Ontology

An ontology is a formal description to capture a shared understanding knowledge of specific domain. It consists of a vocabulary which illustrates all the concepts that are related with the domain and also the existing relations and constraints between them [17]. Moreover, ontology can organize the concepts through classification according to the intended meaning of the vocabulary and put hierarchy on these [18]. Ontologies have become known as an acceptable way to integrate semantic knowledge into raw text (documents). They are conceptual models and their goal is to provide unambiguous knowledge and also support automatic semantic interpretation of textual information [5,6].

The passive design of ontology is based on Information Extraction systems where the goal is to associate an occurring term in the text with a concept including into ontology. However, this domain-based ontology can be populated incrementally using semantic reasoning, deriving facts that are not expressed in ontology explicitly and enrich its knowledge base. Semantic reasoning can specify inference rules using description language and so it can saturate the ontology through inference techniques.

To design and implement the ontology we use the TopBraid Composer(TBC) software. We have interacted with the expert of the Prader-Willi Syndrome to identify the main entities involved in Prader Willi syndrome and how they are related to each other. We have then created a deductive RDF dataset which is a set of RDF facts and rules on top of them. This deductive RDF dataset captures the hierarchy of classes (concepts) in the ontology, their corresponding instances and also the properties (relations) between them. The hierarchical format of ontology is displayed in Figure (3.1), Figure (3.2).

Additionally, we have defined some new classes in order to create high-level conceptual sets enabling to express queries which are more flexible. For example, instead of having to distinguish between the hormone classes and instances, we have built a new class HormoneSet class which includes all the previous information as instances. As a consequence, the user can submit more flexible queries without knowing which related information has been declared as class or as an instance. To achieve this, we have inserted SPARQL Rules (SPIN) in TBC which is a collection of RDF vocabularies to let us define new functions and inferencing rules for the new classes: ProteinSet, AnomalySet, TroubleSet, SymptomSet, GeneSet, NeurotransmitterSet, TreatmentSet, HormoneSet, BrainAnatomySet. For more details, see Annex: [2]- [10].



```
Classes
rdfs:Resource (715)
  owl:Thing (253)
    owl:Nothing
    ◆ pw:ACTH
    ◆ pw:Affabulation
    ● pw:Animal_Models (2)
    ● pw:AnimalModelsSet_PW (3)
    ● pw:Anomaly_PW (6)
        ● pw:Anomaly_Brain_Development_PW
        ● pw:Genetic_Anomaly_PW (6)
            ● pw:AC15P (6)
        ● pw:Hormonale_Deficiency_PW
    ● pw:AnomalySet_PW (11)
    ◆ pw:Boulimia
    ● pw:Brain_Anatomy
        ● pw:Dopaminergic_Circuits
        ● pw:Hypothalamus_Arcuate_Nucleus
        ● pw:VTA
    ● pw:BrainAnatomySet_PW (4)
    ◆ pw:Corticotropin_CRH
    ◆ pw:Craving_for_Food
    ◆ pw:Def_empreinte_pat
    ◆ pw:Deletion_15_pat
    ◆ pw:Deletion_SNORD_116
    ◆ pw:Emotional_Lability
    ◆ pw:Food_Addiction
    ◆ pw:Food_Storage
    ◆ pw:FSH
    ◆ pw:Gastric_Motility_Problems
    ● pw:Gene (7)
    ● pw:GeneSet_PW (8)
    ◆ pw:GH
    ◆ pw:Ghrelin
    ◆ pw:GnRH
    ◆ pw:Hallucination
    ◆ pw:Hip_Displasia
    ◆ pw:Hoarding
    ● pw:Hormone (19)
        ● pw:Adipokine_Hormones (2)
        ● pw:HH (13)
        ● pw:Thyroid_Hormones (2)
    ● pw:HormoneSet_PW (23)
    ◆ pw:Hyponatremia
    ◆ pw:IGF-1
    ◆ pw:IGFBP-3
    ◆ pw:Impulsiveness
    ◆ pw:Infant_Anorexia
    ◆ pw:Insulin
    ◆ pw:LHRH
    ◆ pw:MAGEL2
    ● pw:Neurotransmitter (1)
    ● pw:NeurotransmitterSet_PW (2)
    ◇ pw:Obesity
    ◆ pw:Obsession_with_Food
    ◆ pw:Ophtalmologic_Problems
    ◆ pw:Oxytocin
    ● pw:Paper
    ◆ pw:Prolactin
    ● pw:Protein (3)
    ● pw:ProteinSet_PW (4)
    ◆ pw:Rigidity_of_Mind
    ◆ pw:Satiety_Deficit
    ◆ pw:Scoliosis
    ● pw:Sentence
    ● pw:Set_PW (94)
    ◆ pw:Skin_Picking
    ● pw:Symptom_PW (10)
    ● pw:SymptomSet_PW (11)
    ◆ pw:Translocation_15_pat
    ● pw:Treatment
        ● pw:Fluoxetine_Treatment
        ● pw:GH_Treatment
        ● pw:Modafinil_Treatment
    ● pw:TreatmentSet_PW (4)
    ● pw:Troubles_PW (13)
        ● pw:Behaviour_Troubles_PW (8)
            ● pw:Deficit_of_Social_Skills
            ● pw:Eating_Disorder_PW (7)
                ● pw:Hyperphagia (1)
        ● pw:Cognitive_Deficit_PW
            ● pw:Executive_Dysfuntions
            ● pw:Learning_Problems
        ● pw:Physical_Development_Troubles_PW
        ● pw:Psychiatric_Troubles_PW (5)
            ● pw:Compulsive_Behaviour
            ● pw:Psychosis
        ● pw:Sleep_Troubles_PW
            ● pw:Apneas
            ● pw:Excessive_Day_Sleepiness
    ● pw:TroubleSet_PW (28)
    ◆ pw:TSH
    ◆ pw:UPD
```

The first level of classes in PW indicates with the symbol ● and all the instances from classes and subclasses in PW appears with symbol ◆

The second level of classes (subclasses) of Prader-Willi ontology indicates with the symbol ● and they aligned righter with increasing level of subclasses

The pw: is the prefix of the Prader-Willi ontology, it abbreviates the URIs of the namespace that is used in this model.

The number of instances and subclasses of this specific class (Troubles)

Figure 3. 1: Hierarchy representation of classes and instances in Prader-Willi ontology

Apart from the concepts related to Prader-Willi domain, we have defined classes and properties in order to connect afterwards the ontology with the content of scientific papers. These classes are called: Paper and Sentence that contains all the scientific articles and their corresponding sentences which mention Prader-Willi entities.

We have also defined 7 properties that relate instances of the different classes (Figure 3.3):

- Associated_With: Relate the class symptoms to class troubles.
- Caused_By: Relate the classes Anomaly or Troubles to class Anomaly.
- isAbout: Relate the class Paper to all including classes of Prader-Willi ontology.
- isaSentenceOf: Relate the class Sentence to class Paper.
- mentions: Relate the class Sentence to all including classes of Prader-Willi ontology.
- Stimulates: Relate the class Hormone to class Hormone.
- publishedIn: relate the class Paper to integer number indicating the year of publication.

These properties permit user to ask fine-grained queries over this combined ontology and obtain precise answers.

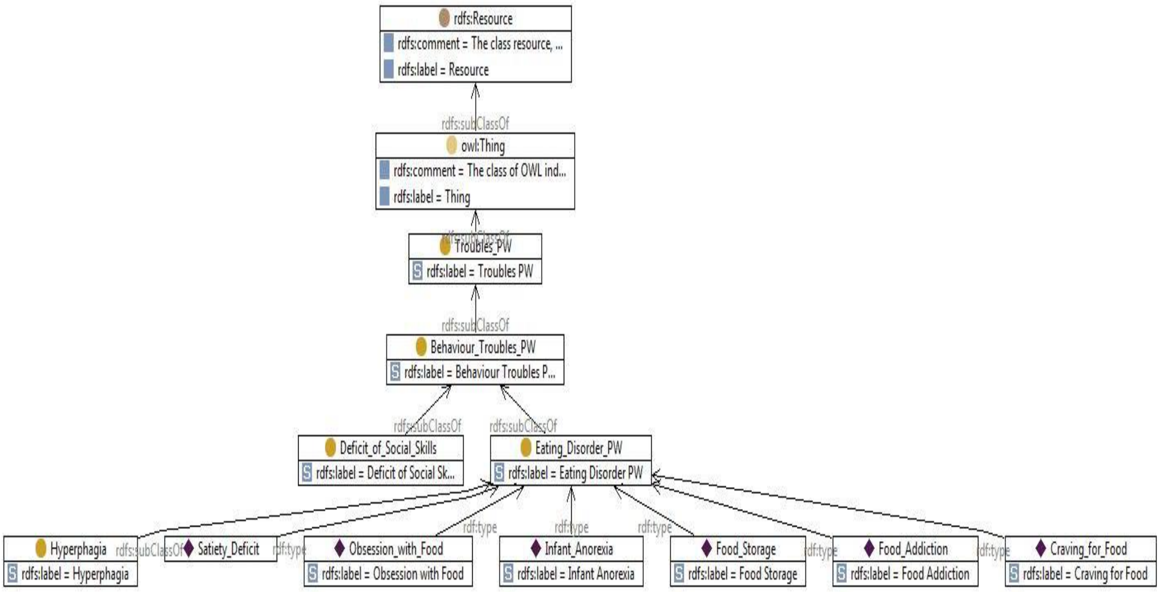
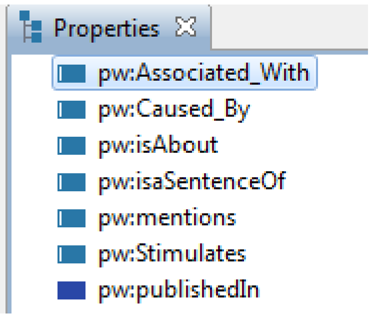


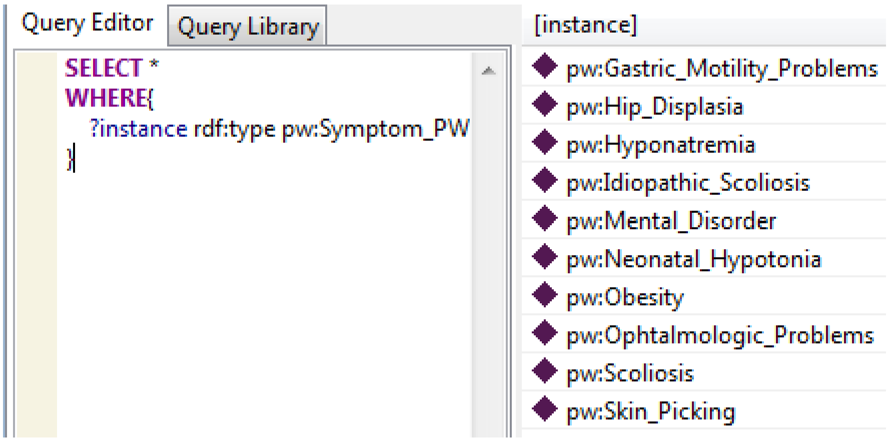
Figure 3. 2: RDF Graph representation of class Troubles


Properties

pw:Associated_With
pw:Caused_By
pw:isAbout
pw:isaSentenceOf
pw:mentions
pw:Stimulates
pw:publishedIn

Figure 3. 3: Prader-Willi properties

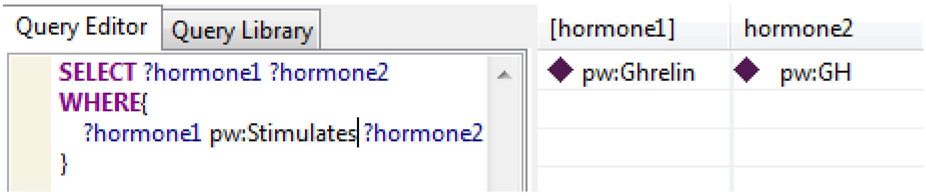
SPARQL is the SQL-like query language for RDF data. The following figures (Figure 3.4, Figure 3.5, Figure 3.6) are screenshots of TopBraid Composer query editor. On left, there is the query. On the rght, its answers returned by TopBraid Composer.



```
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

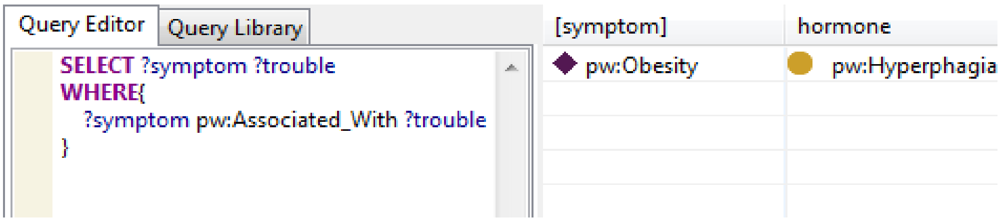
Figure 3. 4: SPARQL query for finding all the instances of symptoms



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

[hormone1]	hormone2
◆ pw:Ghrelin	◆ pw:GH

Figure 3. 5: SPARQL query for finding all the pairs of hormones, stimulating each other.



```
SELECT ?symptom ?trouble
WHERE{
  ?symptom pw:Associated_With ?trouble
}
```

[symptom]	hormone
◆ pw:Obesity	● pw:Hyperphagia

Figure 3. 6: SPARQL query for finding the associations between symptoms and troubles

The current ontology is far from being complete. It would require more interaction with the expert. Nevertheless, the important part is that it can be easily updates by adding new classes, instances or properties using TopBraid Composer software.

4. Prader-Willi Extended to the description of scientific corpus

We have enriched the Prader-Willi ontology by adding new triples in order to associate the scientific papers with Prader-Willi entities. To achieve this mapping, we have implemented a Perl script, responsible for automatic filtering of scientific corpus and to extract sentences related with the PW concepts.

Firstly, we have asked the expert to provide for each formal term of the ontology, the different textual expressions that correspond to them in scientific papers. You can see more details about the formal terms of our ontology and the corresponding textual expressions in the Annex: [1]. The expert gave us all the alternative expressions that correspond to Prader-Willi entities (classes, instances, properties) in scientific corpus.

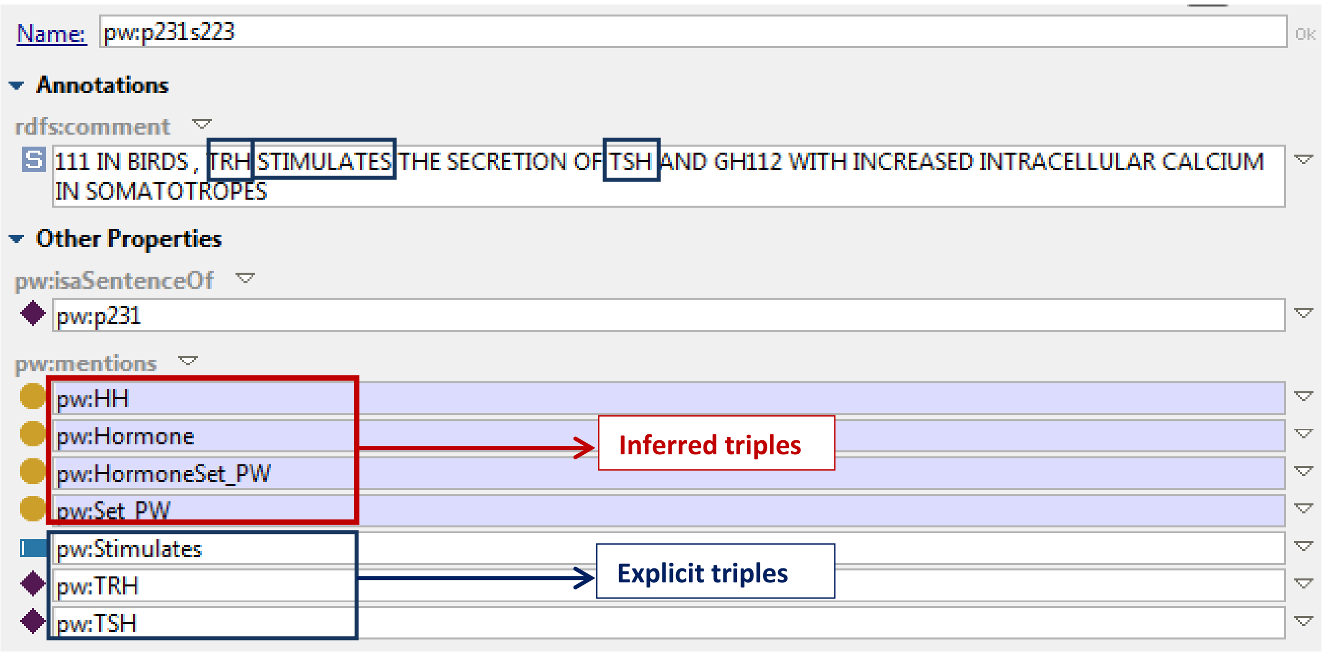
In order to extract meaningful sentences from the paper, we have firstly transformed the pdf format of files into textual representation in order to obtain all the sentences of each article. To identify the sentences to extract, we have kept the sentences that contain at least one textual expression associated to Prader-Willi entities.

Then, we have transformed the information into RDF triples that are inserted into the ontology. More precisely, each paper is assigned to a unique identifier (e.g. p1, p2, p3, etc.), and each sentence too. For instance, the identifier p1s1 describes that the first sentence is extracted from the paper p1. We relate each paper's identifier to the paper's title by using the RDF property rdfs:label, and its year of publication by using the Prader-Willi's property, "publishedIn". Similarly, each sentence identifier is associated to its textual content by using the RDF property rdf:comment and also with which paper it belongs to by using the Prader-Willi's property, "isaSentenceOf".

To associate the extracted sentences with the Prader-Willi entities, we have inserted triples by using the property of our ontology, "mentions". Apart from the explicit triples of property "mentions", we have inserted SPIN inferencing rules in order to saturate our ontology and insert inferred triples. For example, if there is an inserted triple <pw:p1608s14 pw:mentions pw:Oxytocin>, where Oxytocin is an instance of HH(Hypothalamic Hormones) and also HH is a subclass of Hormone, the triples <pw:p1608s14 pw:mentions pw:HH> and <pw:p1608s14 pw:mentions pw:HH> will be inferred.

For this reason, we have inserted two SPIN rules in our ontology in order to populate our ontology by inserting inferred triples. The first one is to take the class to which the instance belongs and the second one to find all the superclasses of the class in which the current instance is included. The two SPIN rules in order are the following:

```
1)  CONSTRUCT {
             ?s pw:mentions ?class .}
    WHERE {
             ?s pw:mentions ?instance .
             ?instance a ?class .
             FILTER (?class != owl:Class) .
             FILTER (?class != owl:Thing) .
             FILTER (?class != rdf:Property) .
             FILTER (?class != owl:ObjectProperty) .}


2)  CONSTRUCT {
             ?s pw:mentions ?superClass .
    }
    WHERE {
             ?s pw:mentions ?class .
              ?class (rdfs:subClassOf)* ?superClass .
              FILTER (?class != owl:Class) .
             FILTER (?class != owl:Thing) .
              FILTER (?class != rdf:Property) .
              FILTER (?class != owl:ObjectProperty) .
              FILTER (?superClass != owl:Class) .
              FILTER (?superClass != owl:Thing) .
              FILTER (?superClass != rdfs:Resource) .}
```



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

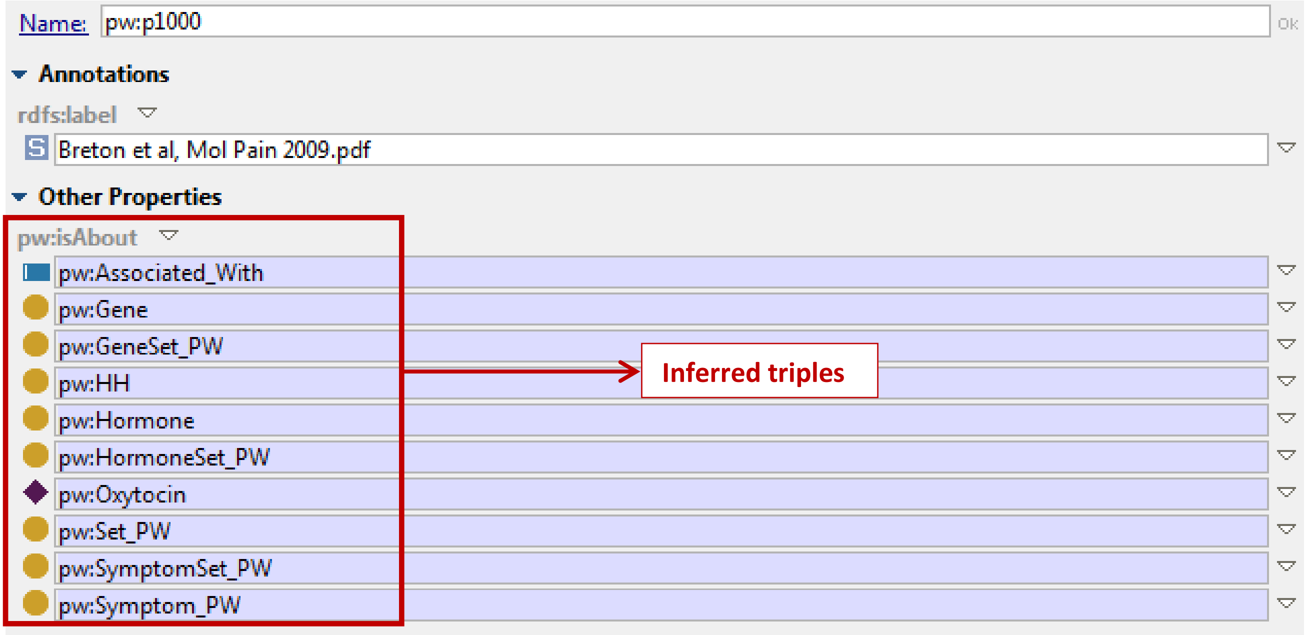
pw:mentions

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

Figure 4. 1: The property mentions of Prader-Willi ontology, showing the explicit triples and also the inferred triples which are created by SPIN rules.

Apart from the sentences associated to the Prader-Willi entities, we used the property "isAbout" from our ontology to associate papers with Prader-Willi concepts. We haven't inserted explicitly triples into our ontology because they could be inferred by SPIN rule. The triples of property "isAbout" will be inserted through a SPIN rule by using the explicit and inferred triples of the property "mentions" and the existing triples of property "isaSentenceOf" which declares the included sentences for each paper. The following SPIN rule is:

```
CONSTRUCT {

   ?p pw:isAbout ?f .
}
WHERE {
   ?s pw:isaSentenceOf ?p .
   ?s pw:mentions ?f .
}
```



Name: pw:p1000

Annotations

rdfs:label

Breton et al, Mol Pain 2009.pdf

Other Properties

pw:isAbout

- pw:Associated_With
- pw:Gene
- pw:GeneSet_PW
- pw:HH
- pw:Hormone
- pw:HormoneSet_PW
- pw:Oxytocin
- pw:Set_PW
- pw:SymptomSet_PW
- pw:Symptom_PW

Inferred triples

Figure 4. 2: The property isAbout of Prader-Willi ontology, showing the inferred triples which are created by SPIN rule.

The size of Prader-Willi domain ontology was 160 KB and after the mapping with scientific corpus, we apply saturation into ontology and its size becomes 264 MB. The number of scientific papers which is utilized as an input into the Perl script is 5950 and their corresponding sentences from the scientific corpus amounts to 359787. Our final ontology consists of: 2941688 RDF triples, 27 classes, 7 properties and 365798 instances.

5. Query-based Analysis and Access

In this section, we can 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 allows us to do simple statistical analysis of corpus, the second category enables fine-grained semantic analysis of its content and the third one shows some queries that can be directly expressed by the expert. The user can type SPARQL queries using TopBraid Composer software and access easily the updated Knowledge Base of the ontology.

5.1. Query-based analysis of corpus

We have made some queries in order to analyze to what extent the corpus is related to the domain ontology. Firstly, we have submitted a query to find the top 50 papers that contain the most sentences and ranked them accordingly (Figure 5.1.1). We can see that our approach extracts up to 1061 Prader-Will related sentences which imply that our limited Prader-Willi vocabulary can return significant amount of information.

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
Kaiya 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

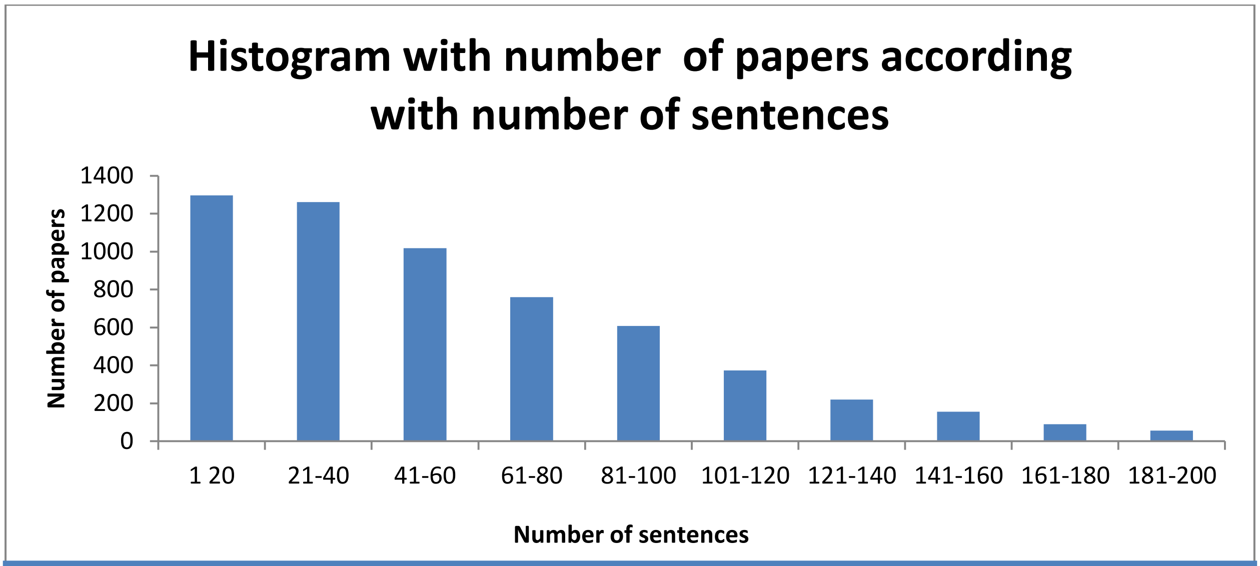
Bodnar et al, Pept 2013.pdf	311
Inui et al, Nature Rev 2001.pdf	299
Hasselbalch et al, Dan Med Bull 2010.pdf	297
Bereket et al, Obes Rev 2012.pdf	296
Koletzo, Int J Obesity 2007 abstract congr?s.pdf	291

Sparql:

```
SELECT ?titlePaper (COUNT(*) AS ?count)

WHERE{

        ?a pw:isaSentenceOf ?s.

        ?s rdfs:label ?titlePaper.
```

Figure 5.1. 1:The top 50 papers ranked by number of Prader-Willi related sentences and its corresponding SPARQL query

Secondly, we compute the distribution among the number of papers and their number of PW-related sentences by submitting a query (Figure 5.1.2). Around 98% of the scientific corpus range from 1 to 100 PW-related sentences and 1.5% of papers has between 101 and 200 sentences. Therefore, the Prader-Willi entities with the corresponding expressions can match approximately more than 100 sentences in average, which means that we have declared enough expressions to extract sentences from our corpus.



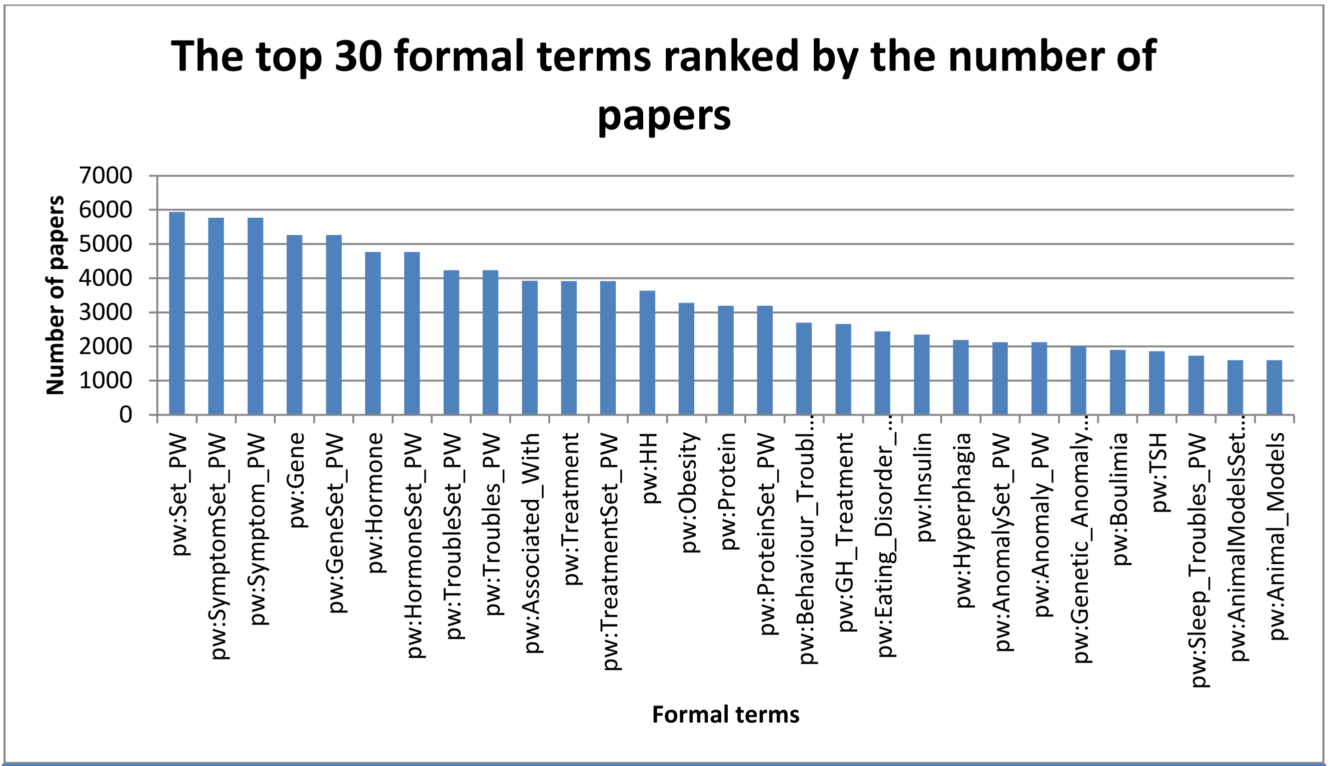
SPARQL:

```
SELECT ?s (COUNT(*) AS ?count)
WHERE{
  ?a pw:isaSentenceOf ?s.}
GROUP BY ?s
ORDER BY DESC(?count)
```

Figure 5.1. 2: Histogram with number of papers according with the number of sentences. Also, the corresponding SPARQL query

Moreover, we have explored the appearance of Prader-Willi entities in the papers. Initially, we have counted the number of papers for each PW concept via SPARQL query (Figure 5.1.3).

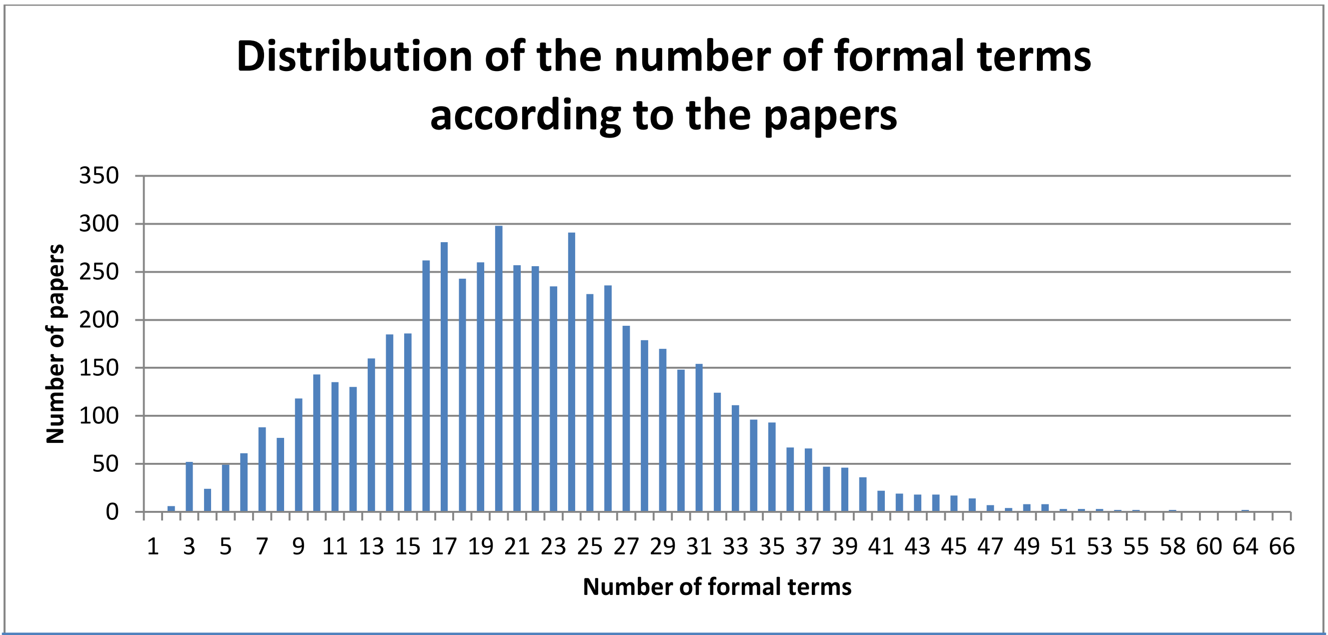
From the Figure 5.1.3, it's noticeable that the most frequent formal terms that are mentioned in the papers are the general entities (Symptom, Gene, Hormone, Troubles), it is because they include all the instances. Also, Obesity appears frequently among the specific PW concepts (instances) and hence we can directly conclude, it is such, because the relation between Obesity and Prader-Willi syndrome is very strong and it's a very good result because it also validates this link from the expert.



```
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
```

Figure 5.1. 3:Top 30 Prader-Willi concepts ranking by the number of papers, that they are mentioned

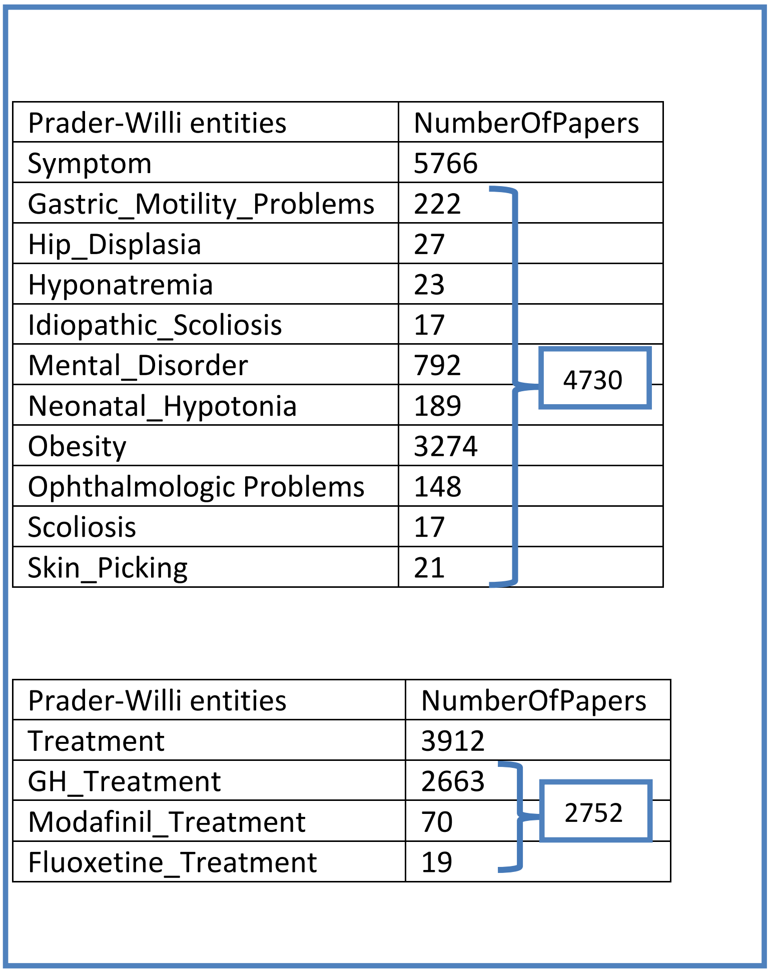
Afterwards, we have investigated the number of papers according to the number of PW entities via a query (Figure 5.1.4). The important remark we can make is that a massive amount of papers include 16-26 PW concepts, which means that our Prader-Willi vocabulary, even if limited can produce satisfactory results.



```
SPARQL:
SELECT ?numFormal_terms COUNT(?numFormal_terms)
WHERE{
SELECT  ?paper ?numFormal_terms
WHERE{
SELECT (COUNT(*) AS ?numFormal_terms) ?paper
WHERE{
 ?paper pw:isAbout ?formalTerm.}
GROUP BY ?paper}
}
GROUP BY ?numFormal_terms
ORDER BY ASC(?numFormal_terms)
```

Figure 5.1. 4:Distribution of the number of formal terms according the papers and its corresponding SPARQL query

Furthermore, we have tried to investigate whether the general concepts include fully or partial redundant information. As we can see, from the tables in Figure 5.1.5, we have concluded that there is a partial redundant information between the general Prader-Willi classes and their including instances. This happens because we have taken into account different expressions for the general Prader-Willi entities and also we have counted many times the instances of the general classes. For instance, you can see the class Symptom and the class Treatment from our ontology has bigger number of references into the papers than the total sum of the papers that contain the instances of them (Figure 5.1.5).



Prader-Willi entities	NumberOfPapers
Symptom	5766
Gastric_Motility_Problems	222
Hip_Displasia	27
Hyponatremia	23
Idiopathic_Scoliosis	17
Mental_Disorder	792
Neonatal_Hypotonia	189
Obesity	3274
Ophthalmologic Problems	148
Scoliosis	17
Skin_Picking	21

Prader-Willi entities	NumberOfPapers
Treatment	3912
GH_Treatment	2663
Modafinil_Treatment	70
Fluoxetine_Treatment	19

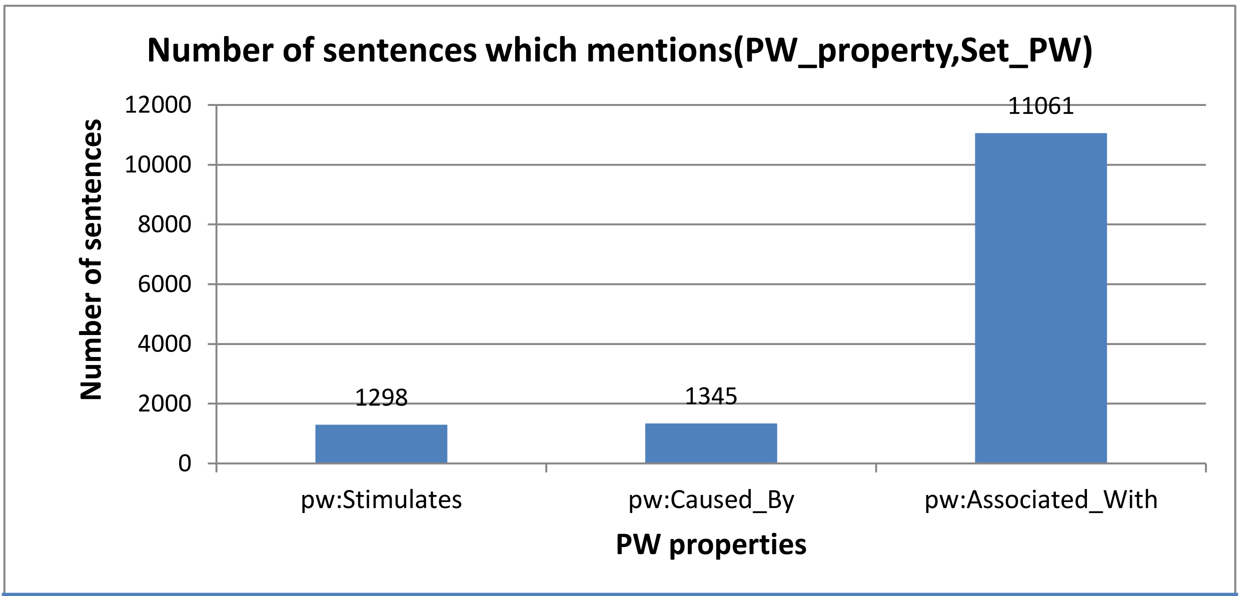
Figure 5.1. 5: Partial redundancy between general Prader-Willi entities and their including instances

5.2. Query-based fine-grained semantic analysis on the corpus

Our final ontology has kept essential information because it has contained the extracted data for each scientific paper which is mapped by vocabulary of Prader-Willi domain ontology. Hence, we can process this extracted data to explore new facts either by investigating co-occurrences between Prader-Willi entities or investigate recent trends of PW concepts during the publication period. In order to make this investigation, we carry out some SPARQL queries.

For the sake of investigation of the co-occurrences between Prader Willi concepts, we have tried to explore this phenomenon by grouping the Prader-Willi entities. Then, we have studied the number of these appearances.

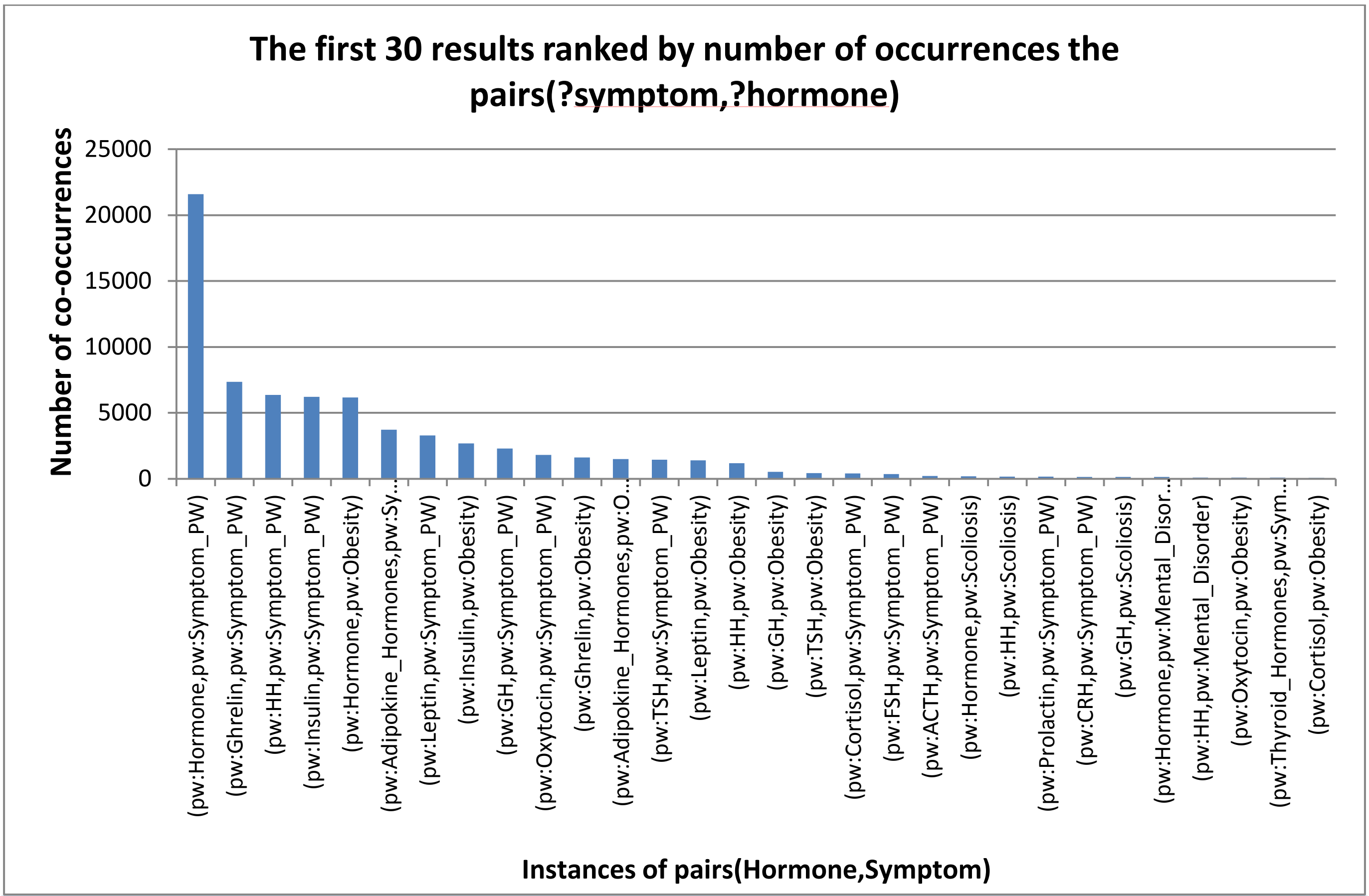
Firstly, we have examined the number of times the Prader-Willi properties' and PW entities' appears in the same sentence by typing a query (Figure 5.2.1).As we can observe from the graph, the most frequent occurrences of Prader-Willi sentences is Associated_With. This relation might be conjoined Prader_Willi concepts and it needs to explore these sentences which contain this relation in order to extract new facts.



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

Figure 5.2. 1: The number of sentences which mentions simultaneously property and any other concept of Prader-WIlli. The cooresponding SPARQL query.

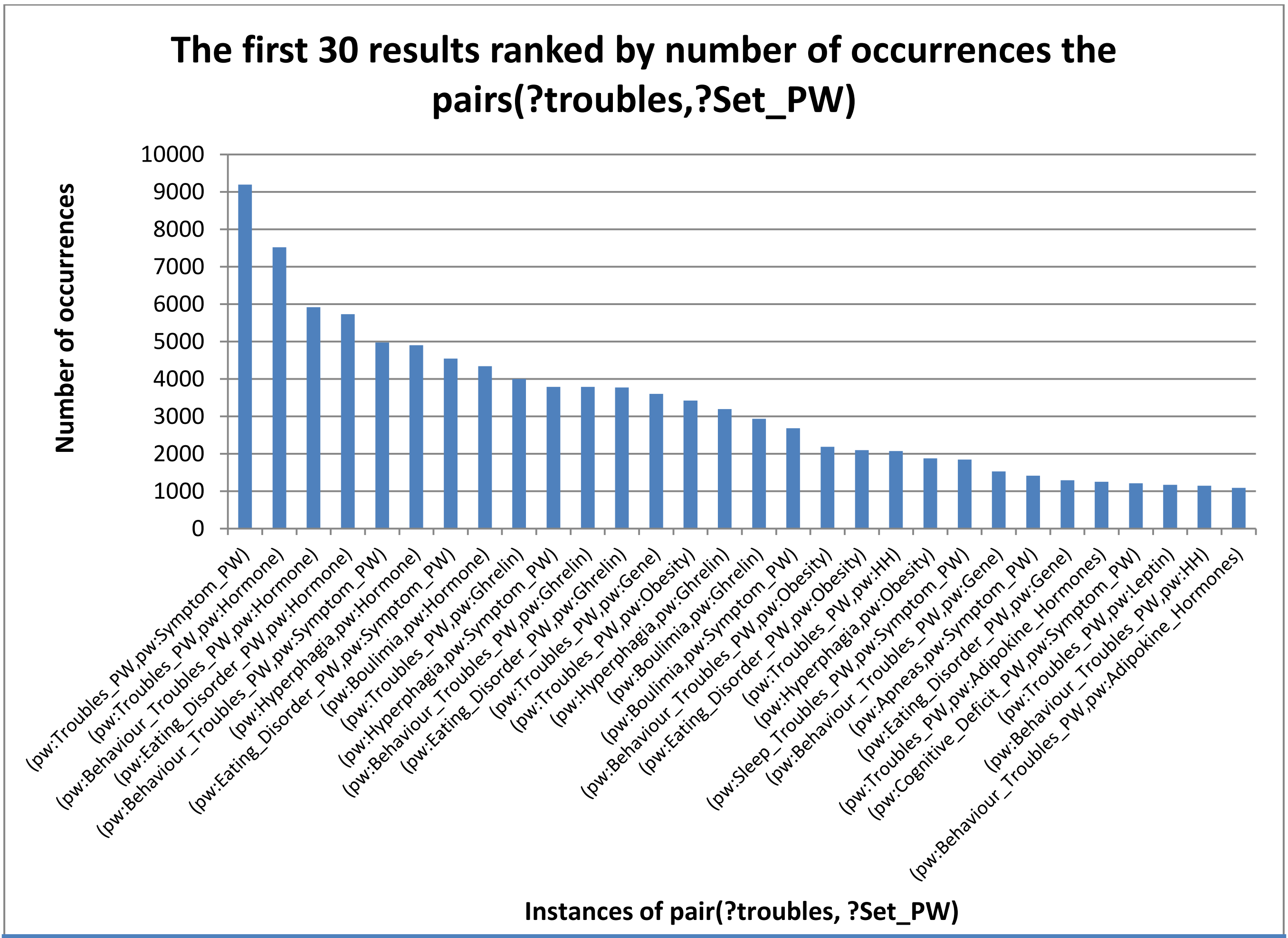
Moreover, we have found the first 30 co-occurrences of symptoms and hormones, which are classed by their frequency (Figure 5.2.2). As we can see from the chart, the number of co-occurrences between the entities Hormone and Symptom reaches approximately 21000, which means that the link between them is robust and strong. Also, the most closed link among Hormone and Symptom is: Leptin with Obesity. So, it might be a good remark to say that Leptin somehow caused Obesity and also we can see from before it’s strongly related with Prader-Willi syndrome.



```
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
```

Figure 5.2. 2: The first 30 co-occurrences,symptoms and hormones,which are ranked by their frequency. The corresponding SPARQL query.

Additionally, we have inquired about the concurrency among the troubles and the remaining Prader-Willi concepts. After, we select the top 30 results with higher concurrency frequency using a query (Figure 5.2.3). From Figure 5.2.3, it’s undeniable that the relationship between troubles and other Prader-Willi concepts is potent, because of the number of sentences reaches up more than 9000. Also, we observe that the most frequent specific pair is Hyperphagia as trouble and Obesity as symptom of Prader-Willi.



```
SPARQL:
SELECT (COUNT(*) AS ?numberOfOccurrence)?trouble ?all
WHERE{
?trouble rdf:type pw:TroubleSet_PW.
?all rdf:type pw:Set_PW.
?s pw:mentions ?trouble.
?s pw:mentions ?all.
MINUS { ?all rdf:type pw:TroubleSet_PW }}
GROUP BY ?trouble ?all
```

Figure 5.2. 3: The first 30 co-occurrences among troubles and any other concept of Prader-Willi ontology,ranked by their number.The corresponding query.

We have typed a query in order to find the co-occurrences between the troubles and the symptoms (Figure 5.2.4). Moreover, from Figure 5.2.4, the symptoms are strongly related with hormones because above 20000 sentences are included in these concepts at the same time. The most specific pair of instances Symptom and Hormone is the Obesity with Insulin.

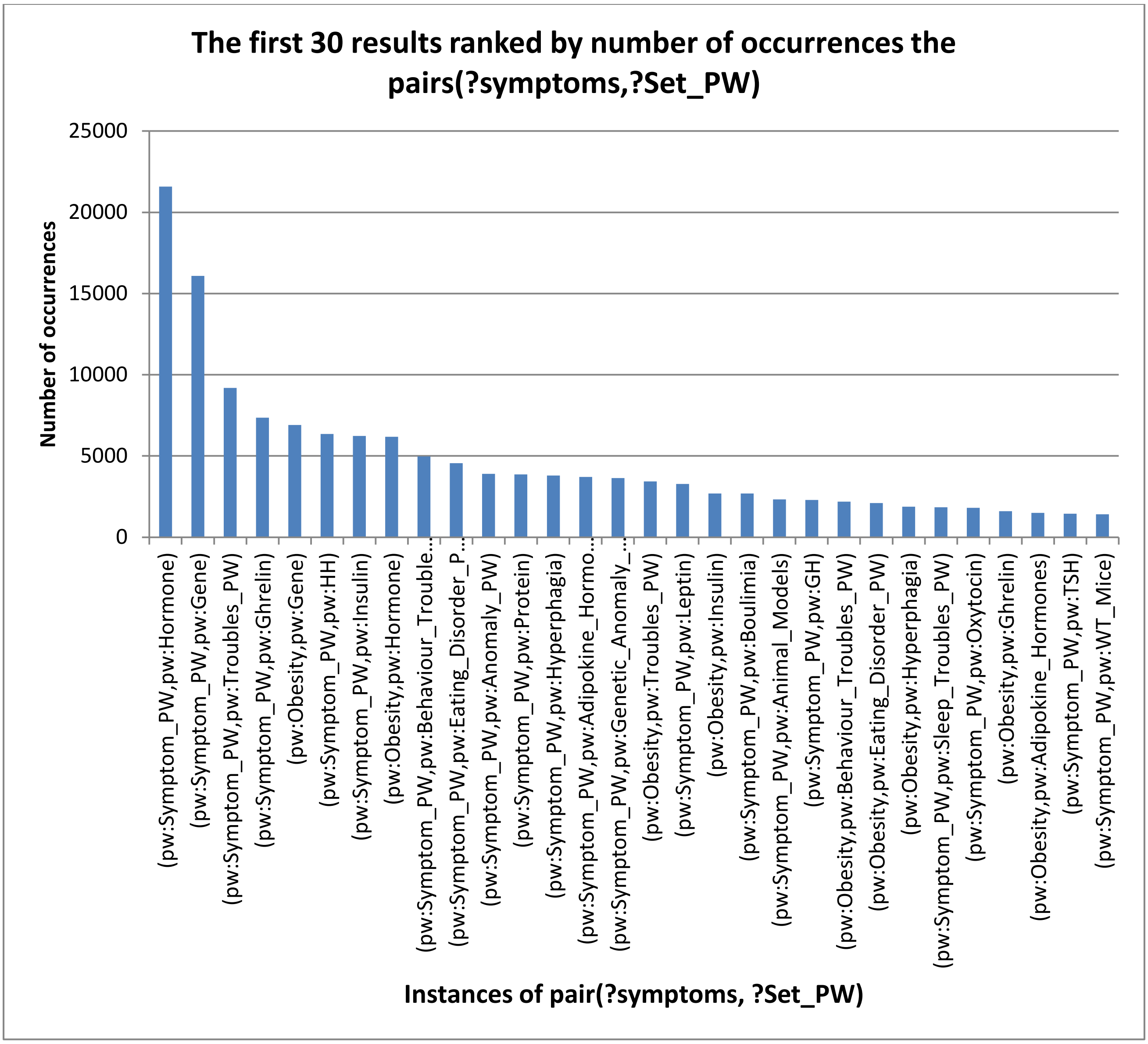
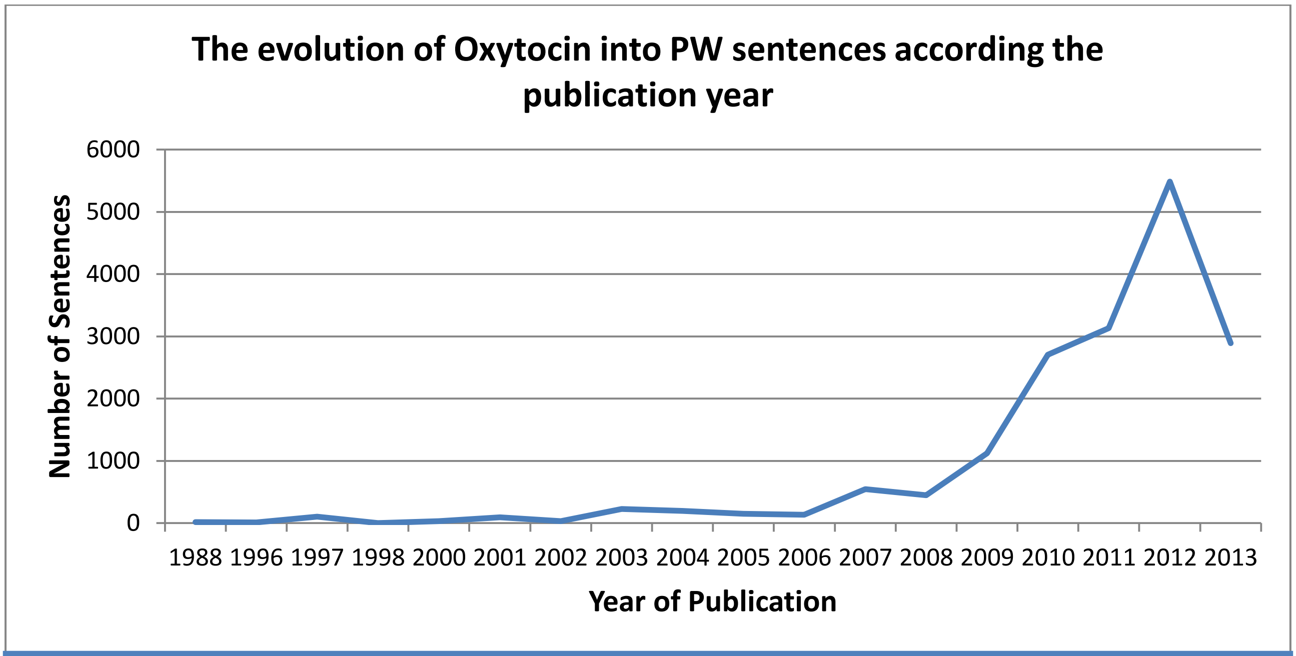


Figure 5.2. 4:The first 30 co-occurrences among symptoms and any other concept of Prader-Willi ontology,ranked by their number. The corresponding SPARQL query.

Notwithstanding, the co-occurrences between the Prader-Willi terms, it was quite interesting to inspect potential trends related to the time-period of the scientific publications. So, we have chosen the Oxytocin which is a hormone and we analyze the evolution of this concept.

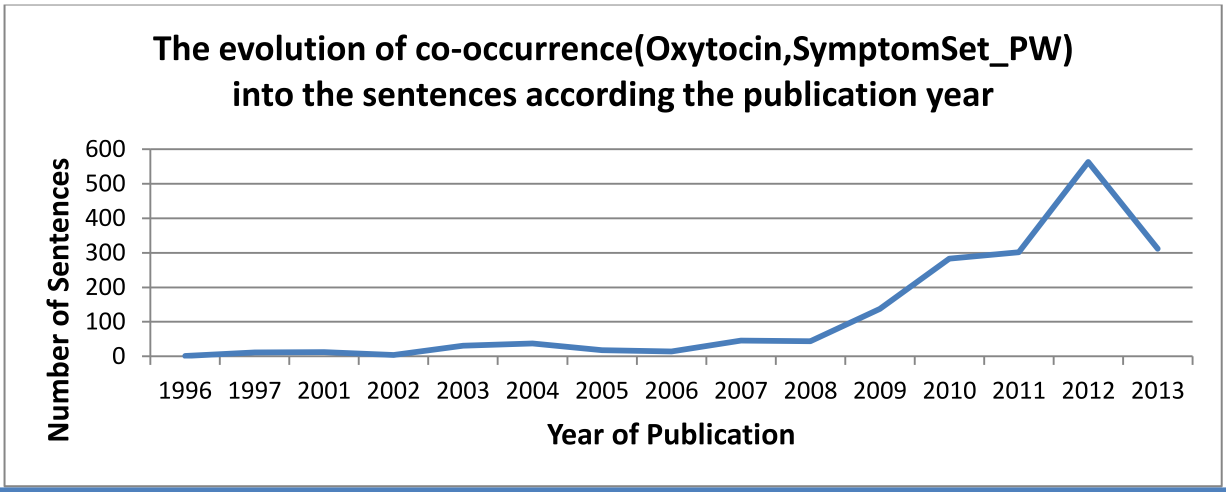
Primarily, we have found the number of sentences that mention the term Oxytocin which is classified by the year of publication via a SPARQL query (Figure 5.2.4). Afterwards, we do survey of co-occurrence of the pairs: Oxytocin and SymptomSet (Figure 5.2.5), Oxytocin and GeneSet (Figure 5.2.6) and also the pair Oxytocin and TreatmentSet during the period of publications (Figure 5.2.7).

For all the graphs on Figures 5.2.4-5.2.8, we have concluded that the term Oxytocin has developed significant tie with Prader-Willi syndrome, the growth mainly has started in 2009. The reason that in 2013 we have had an apparent decline is because of the small amount of papers which are included in our dataset (Figure 5.2.8).



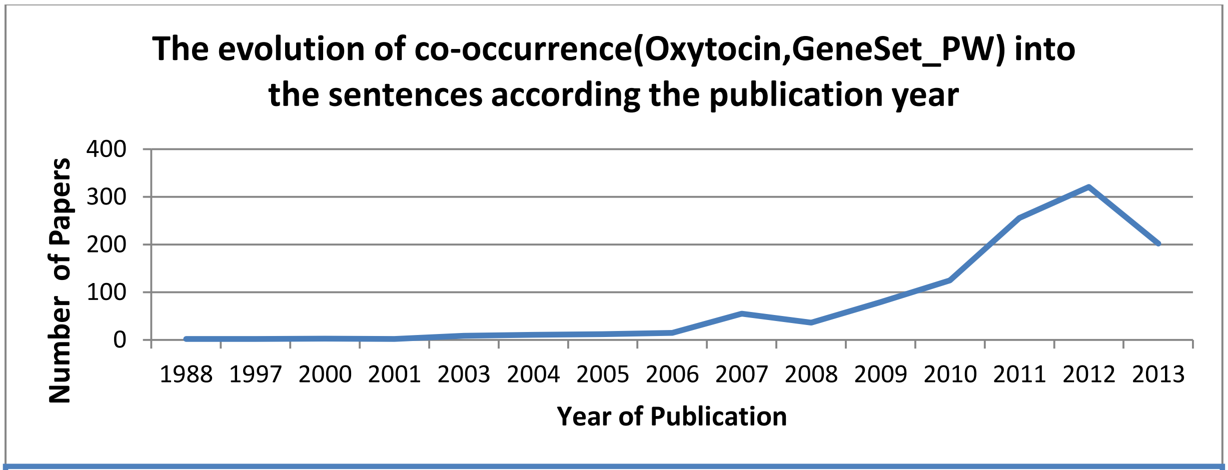
```
SPARQL:
SELECT  (COUNT(?paper) AS ?numOfPapers) ?yearOfPublication
WHERE{
?sentence pw:mentions pw:Oxytocin.
?sentence pw:isaSentenceOf ?paper.
?paper pw:publishedIn ?yearOfPublication.
}
GROUP BY ?yearOfPublication
ORDER BY ASC(?yearOfPublication)
```

Figure 5.2. 5: The number of sentences that mention the term Oxytocin and which is classified by the year of publication. The corresponding query.



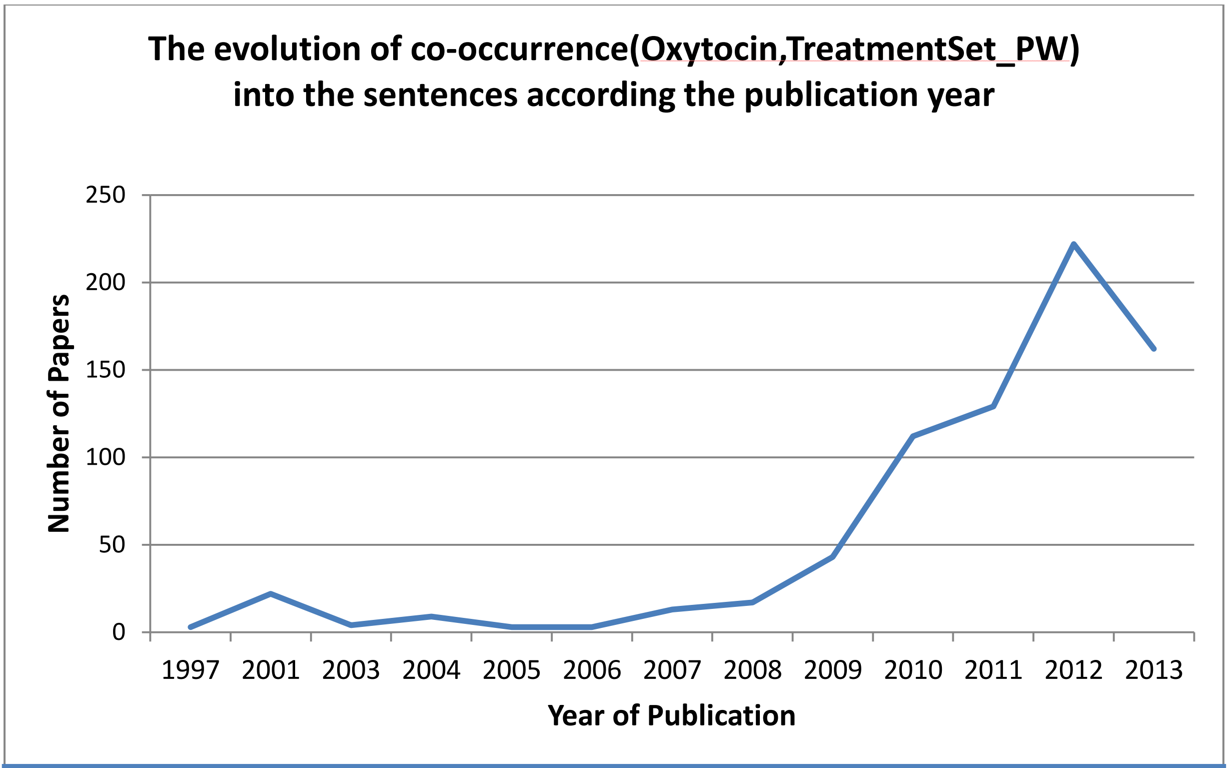
```
SPARQL:
SELECT  (COUNT(?paper) AS ?numOfPapers) ?yearOfPublication
WHERE{
?sentence pw:mentions pw:Oxytocin.
?sentence pw:mentions pw:SymptomSet_PW.
?sentence pw:isaSentenceOf ?paper.
?paper pw:publishedIn ?yearOfPublication.
}
```

Figure 5.2. 6: Survey the co-occurrences the pairs: Oxytocin and SymptomSet. The corresponding SPARQL query.



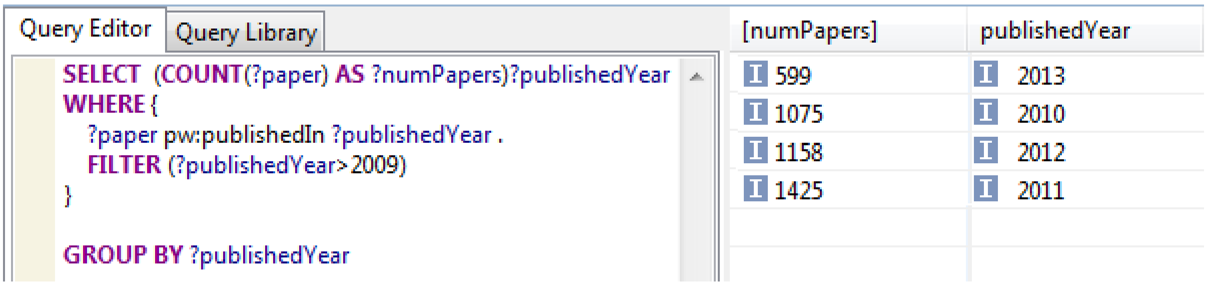
```
SPARQL:
SELECT  (COUNT(?paper) AS ?numOfPapers) ?yearOfPublication
WHERE{
?sentence pw:mentions pw:Oxytocin.
?sentence pw:mentions pw:GeneSet_PW.
?sentence pw:isaSentenceOf ?paper.
?paper pw:publishedIn ?yearOfPublication.
}
GROUP BY ?yearOfPublication
```

Figure 5.2. 7: Survey the co-occurrence of the pairs Oxytocin and GeneSet. The corresponding SPARQL query.



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

Figure 5.2. 8: Survey the co-occurrences of tha pairs: Oxytocin and TreatmentSe. The corresponding SPARQL query.



```
SELECT  (COUNT(?paper) AS ?numPapers)?publishedYear
WHERE{
   ?paper pw:publishedIn ?publishedYear .
   FILTER (?publishedYear> 2009)
}

GROUP BY ?publishedYear
```

[numPapers]	publishedYear
599	2013
1075	2010
1158	2012
1425	2011

Figure 5.2. 9: The number of papers according to the year of publication. The corresponding SPARQL query.

5.3. Interesting queries for the expert

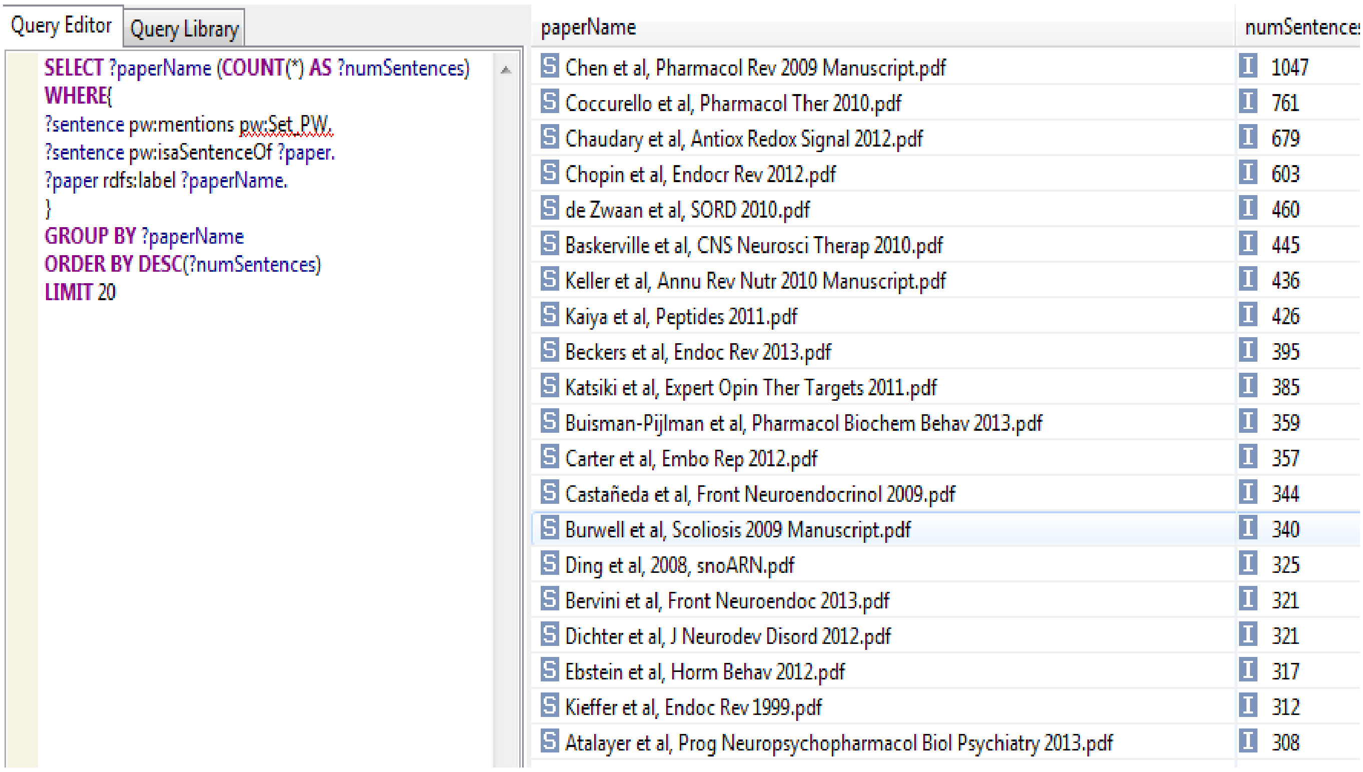
The main purpose of this project is to assist the expert to search interesting information within the bibliographical articles by submitting precise. For this reason, we have made some examples to show what kind of queries user can carry out and receive useful information from the scientific corpus.

The first query has calculated the first 20 papers according to the number of sentences that are related with Prader-Willi in the corpus (Figure 5.3.1).

The second query has found all the papers which mention hormones and symptoms in their sentences and show the content of each sentence (Figure 5.3.2).

The third query, has made more specific the previous query by replacing the hormones with a specific hormone, Dopamine. Also, apart from the symptoms, it has tried to find also the property PW relation, Associated_With that are mentioned in the same sentence and displaying them to the expert (Figure 5.3.3).

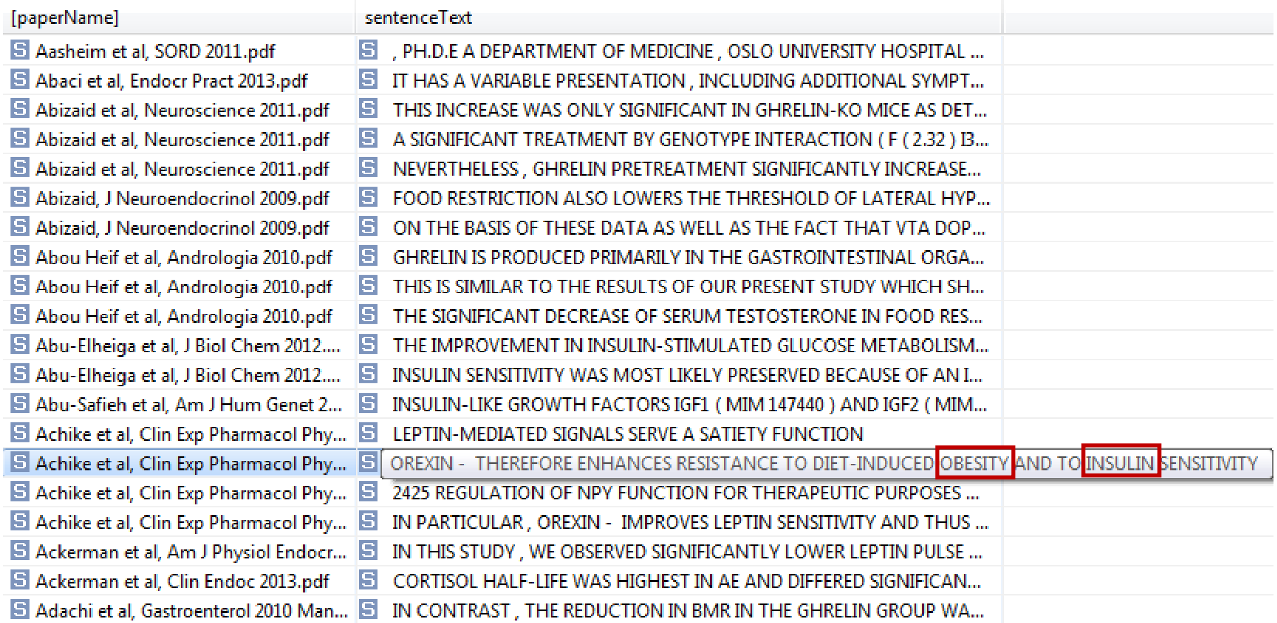
The last query, searches the papers which includes sentences that mention Prader-Willi's anomalies, Prader-Willi's troubles and also to find all the causality links between them by searching the property Caused_By inside the sentences (Figure 5.3.4).



```
SELECT ?paperName (COUNT(*) AS ?numSentences)
WHERE{
?sentence pw:mentions pw:Set_PW.
?sentence pw:isaSentenceOf ?paper.
?paper rdfs:label ?paperName.
}
GROUP BY ?paperName
ORDER BY DESC(?numSentences)
LIMIT 20
```

paperName	numSentence
Chen et al, Pharmacol Rev 2009 Manuscript.pdf	1047
Coccurello et al, Pharmacol Ther 2010.pdf	761
Chaudary et al, Antiox Redox Signal 2012.pdf	679
Chopin et al, Endocr Rev 2012.pdf	603
de Zwaan et al, SORD 2010.pdf	460
Baskerville et al, CNS Neurosci Therap 2010.pdf	445
Keller et al, Annu Rev Nutr 2010 Manuscript.pdf	436
Kaiya et al, Peptides 2011.pdf	426
Beckers et al, Endoc Rev 2013.pdf	395
Katsiki et al, Expert Opin Ther Targets 2011.pdf	385
Buisman-Pijlman et al, Pharmacol Biochem Behav 2013.pdf	359
Carter et al, Embo Rep 2012.pdf	357
Castañeda et al, Front Neuroendocrinol 2009.pdf	344
Burwell et al, Scoliosis 2009 Manuscript.pdf	340
Ding et al, 2008, snoARN.pdf	325
Bervini et al, Front Neuroendoc 2013.pdf	321
Dichter et al, J Neurodev Disord 2012.pdf	321
Ebstein et al, Horm Behav 2012.pdf	317
Kieffer et al, Endoc Rev 1999.pdf	312
Atalayer et al, Prog Neuropsychopharmacol Biol Psychiatry 2013.pdf	308

Figure 5.3. 1: The first 20 papers according to the number of PW-related sentences and the corresponding SPARQL query.

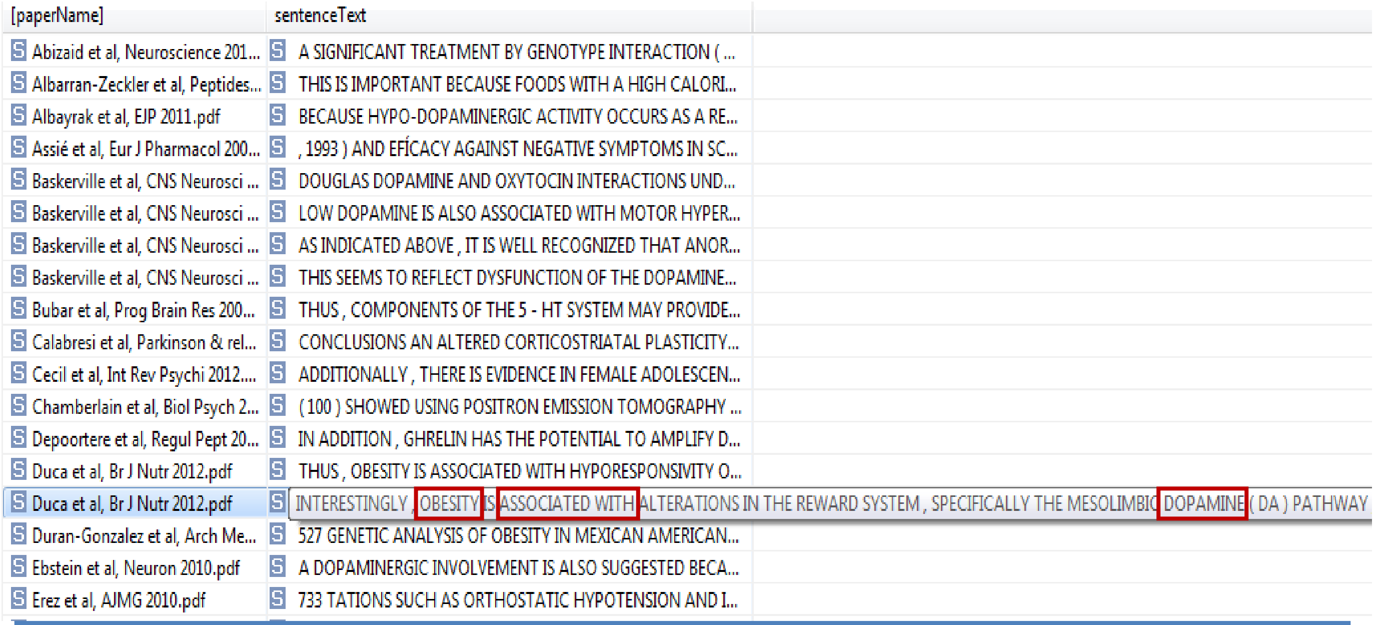


[paperName]	sentenceText
Aasheim et al, SORD 2011.pdf	, PH.D.E A DEPARTMENT OF MEDICINE , OSLO UNIVERSITY HOSPITAL ...
Abaci et al, Endocr Pract 2013.pdf	IT HAS A VARIABLE PRESENTATION , INCLUDING ADDITIONAL SYMPT...
Abizaid et al, Neuroscience 2011.pdf	THIS INCREASE WAS ONLY SIGNIFICANT IN GHRELIN-KO MICE AS DET...
Abizaid et al, Neuroscience 2011.pdf	A SIGNIFICANT TREATMENT BY GENOTYPE INTERACTION (F (2.32) 13...
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.}
```

Figure 5.3. 2: All the papers which mention Hormones and Symptoms and displaying the textual content of these sentences. The corresponding SPARQL query.

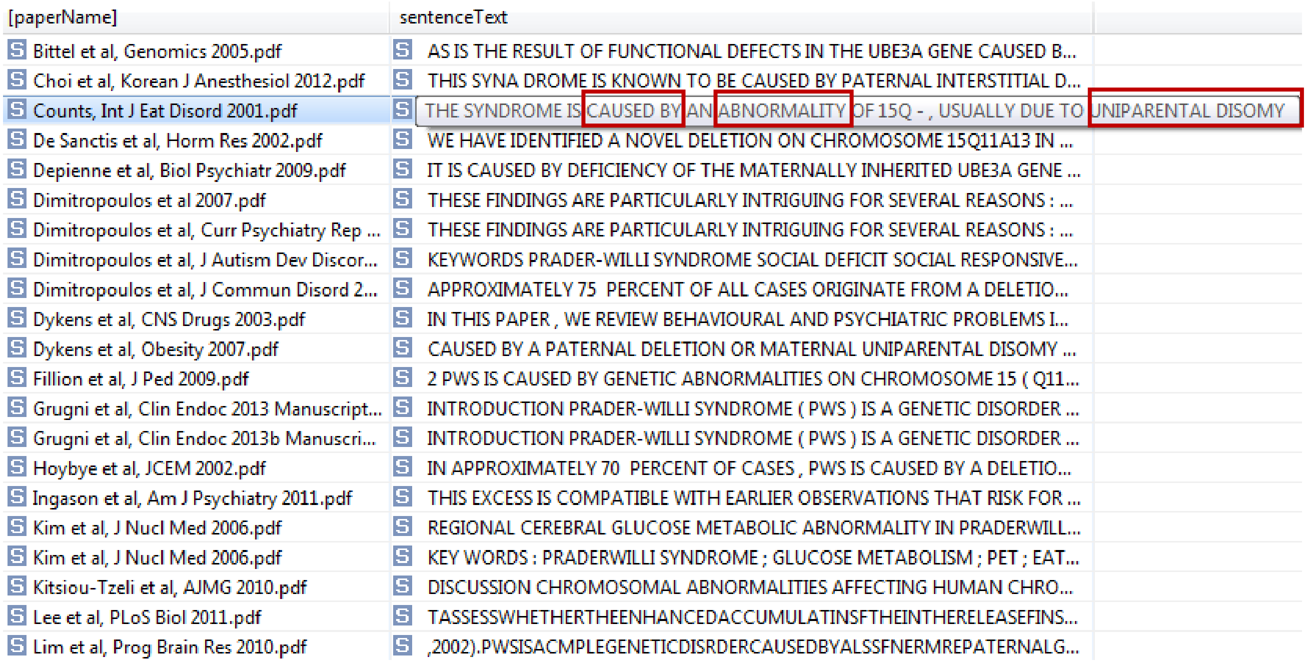


[paperName]	sentenceText
Abizaid et al, Neuroscience 201...	A SIGNIFICANT TREATMENT BY GENOTYPE INTERACTION (...
Albarran-Zeckler et al, Peptides...	THIS IS IMPORTANT BECAUSE FOODS WITH A HIGH CALORI...
Albayrak et al, EJP 2011.pdf	BECAUSE HYPO-DOPAMINERGIC ACTIVITY OCCURS AS A RE...
Assié et al, Eur 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) PATHWAY
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.
}
```

Figure 5.3. 3: All the papers which mention Symptoms and the relation Associated_With and displaying the textual content of these sentences.The corresponding SPARQL query.



[paperName]	sentenceText
Bittel et al, Genomics 2005.pdf	AS IS THE RESULT OF FUNCTIONAL DEFECTS IN THE UBE3A GENE CAUSED B...
Choi et al, Korean J Anesthesiol 2012.pdf	THIS SYNA DROME IS KNOWN TO BE CAUSED BY PATERNAL INTERSTITIAL D...
Counts, Int J Eat Disord 2001.pdf	THE SYNDROME IS CAUSED BY AN ABNORMALITY OF 15Q - , USUALLY DUE TO UNIPARENTAL DISOMY
De Sanctis et al, Horm Res 2002.pdf	WE HAVE IDENTIFIED A NOVEL DELETION ON CHROMOSOME 15Q11A13 IN ...
Depienne et al, Biol Psychiatr 2009.pdf	IT IS CAUSED BY DEFICIENCY OF THE MATERNALLY INHERITED UBE3A GENE ...
Dimitropoulos et al 2007.pdf	THESE FINDINGS ARE PARTICULARLY INTRIGUING FOR SEVERAL REASONS : ...
Dimitropoulos et al, Curr Psychiatry Rep ...	THESE FINDINGS ARE PARTICULARLY INTRIGUING FOR SEVERAL REASONS : ...
Dimitropoulos et al, J Autism Dev Discor...	KEYWORDS PRADER-WILLI SYNDROME SOCIAL DEFICIT SOCIAL RESPONSIVE...
Dimitropoulos et al, J Commun Disord 2...	APPROXIMATELY 75 PERCENT OF ALL CASES ORIGINATE FROM A DELETIO...
Dykens et al, CNS Drugs 2003.pdf	IN THIS PAPER , WE REVIEW BEHAVIOURAL AND PSYCHIATRIC PROBLEMS I...
Dykens et al, Obesity 2007.pdf	CAUSED BY A PATERNAL DELETION OR MATERNAL UNIPARENTAL DISOMY ...
Fillion et al, J Ped 2009.pdf	2 PWS IS CAUSED BY GENETIC ABNORMALITIES ON CHROMOSOME 15 (Q11...
Grugni et al, Clin Endoc 2013 Manuscript...	INTRODUCTION PRADER-WILLI SYNDROME (PWS) IS A GENETIC DISORDER ...
Grugni et al, Clin Endoc 2013b Manuscri...	INTRODUCTION PRADER-WILLI SYNDROME (PWS) IS A GENETIC DISORDER ...
Hoybye et al, JCEM 2002.pdf	IN APPROXIMATELY 70 PERCENT OF CASES , PWS IS CAUSED BY A DELETIO...
Ingason et al, Am J Psychiatry 2011.pdf	THIS EXCESS IS COMPATIBLE WITH EARLIER OBSERVATIONS THAT RISK FOR ...
Kim et al, J Nucl Med 2006.pdf	REGIONAL CEREBRAL GLUCOSE METABOLIC ABNORMALITY IN PRADERWILL...
Kim et al, J Nucl Med 2006.pdf	KEY WORDS : PRADERWILLI SYNDROME ; GLUCOSE METABOLISM ; PET ; EAT...
Kitsiou-Tzeli et al, AJMG 2010.pdf	DISCUSSION CHROMOSOMAL ABNORMALITIES AFFECTING HUMAN CHRO...
Lee et al, PLoS Biol 2011.pdf	TASSESSWHETHERTHEENHANCEDACCUMULATINSFTHEINTHERELEASEFINS...
Lim et al, Prog Brain Res 2010.pdf	,2002).PWSISACMPLEGENETICDISRDERCAUSEDBYALSSFNERMREPATERNALG...

SPARQL:

```
SELECT ?paperName ?sentenceText
WHERE{
?sentence pw:mentions pw:Caused_By.
?sentence pw:mentions pw:AnomalySet_PW.
?sentence pw:mentions pw:TroubleSet_PW.
?sentence rdfs:comment ?sentenceText.
?sentence pw:isaSentenceOf ?paper.
?paper rdfs:label ?paperName.
}
```

Figure 5.3. 4: All the papers which mention Anomalies, Troubles and the relation Caused By and displaying the textual content of these sentences. The corresponding SPARQL query.

6. Conclusion and Perspectives

In this project, we addressed the problem rising from the increasing number of bibliographical articles, in specific the literature of medicine. The domain expert faces difficulties to keep herself up to date with the current state of research. The result is that information is stored in form of plain text, hence it requires time by the expert to read it and to extract the essential information by herself. Although, some applications and search engines where user can submit queries, already exist for this aim, in general they are not efficient. The reason is that the expert has some constraints to express her queries as it's only keyword-based.

However, our approach counts a solution to overcome the problems as it gives the opportunity to the expert to express her interest with more precise queries beyond simple keywords. Therefore, it helps the expert to receive specific information according to her preferences in an easy and quick way.

Specifically we have designed a novel approach by creating a domain ontology which associates Prader-Willi concepts to scientific articles. The ontology contains all the Prader-Willi related sentences which are derived from the corpus. Our approach identifies the Prader-Willi concepts from each sentence and consequently declares which concepts are mentioned for each paper. All this information has contributed to populate the ontology by increasing the knowledge base of the domain of interest.

Using TopBraid Composer software, the expert can easily submit complex queries by using SPARQL query language. SPARQL is the most powerful way to express queries comparing to traditional keyword based tools. Our method can return a set of relevant results which are included as semantic relations into our ontology. It can help the expert to easily find fine-grained information within a textual corpus by displaying the textual content of the sentences in the query's results.

Also, we have shown through SPARQL queries that it's an efficient way to probe the declarative knowledge-based approach for computing variety of statistics on the corpus in order to get some useful facts.

Moreover, with our approach we can get potentially significant outcomes. Firstly, we can inspect the trend of Prader-Willi concepts during the publication period and we can observe the current evolution among them in the research. Secondly, the important thing is to quickly observe which Prader-Willi concepts are contained in each paper and extract the meaningful sentences related to Prader-Willi entities. Therefore, user can easily recognize which concepts of Prader-Willi are strongly related with each other. For instance, the expert can be certain which hormones and which symptoms are strongly related with each other by examining the number of their co-occurrences from the extracted information. Finally, it's easy to define which instances of our ontology are more frequent in our corpus and we can conclude some significant results.

Even though our ontology contains a small amount of Prader-Willi entities, we have gained a satisfactory number of sentences related to Prader-Willi and we have explored them in order to deduce meaningful conclusions. Also, if we want to update our ontology with either new Prader-Willi concepts (entities, relations, instances) or expressions associated to Prader-Willi ontology's vocabulary, it's straightforward. We can add this new information in our ontology automatically either by asking the expert to give us the data or semantically by using Inferencing SPIN rules. Also, we don't need to make changes neither in our script nor in our SPARQL queries for providing statistics and analyzing the extracted information. The only thing that might need changes is the set of expressions which corresponds to our ontology in order to detect inside the corpus.

One possible direction of our future work is to improve our approach by incorporating a dictionary such as Thesaurus to obtain all the compound terms or acronyms of our topic. Another potential improvement is that we can extract new facts by exploring more exhaustively the extracted information, especially, the relation "isAssociatedWith", which is the most frequent property in our ontology. We could detect new concepts in the sentences which contain the relation "isAssociatedWith" and to update the ontology's knowledge base. This problem can be addressed by Natural Language Processing tools (NLP).

NLP can serve many tasks into the raw text. For example, it provides named entity recognition (NER) in order to determine which items in the text map to proper Prader-Willi concepts and to determine the new entities in our ontology. Also, it can make morphological segmentation in order to take into account all the morphological variations of the words and expressions that we are interested. For instance, the formal term Treatment can simply model all possible forms of a word (treat, treatments, treating). Another potential interesting NLP subtask is to make relationship extraction, meaning to extract new relations among the Prader-Willi concepts.

7. BIBLIOGRAPHY

1. Malo, Pekka, et al. “Semantic content filtering with Wikipedia and ontologies.” Data Mining Workshops (ICDMW), 2010 IEEE International Conference on. IEEE, 2010.
2. Mori, Matsuo, et al. “Finding user semantics on the web using word co-occurrence information”, In Proc. Int’l. Workshop on Personalization on the Semantic Web, 2005
3. Perez-Iratxeta, Carolina, et al. “Update on XplorMed: a web server for exploring scientific literature.” Nucleic Acids Research 31. 13 (2003): 3866-3868.
4. Cassidy,Driscoll.”Prader-Willi syndrome.”European Journal of Human Genetics (2009) 17:3–13
5. Malo, Pekka, et al. "Semantic content filtering with Wikipedia and ontologies." Data Mining Workshops (ICDMW), 2010 IEEE International Conference on. IEEE, 2010.
6. Spasic, Irena, et al. "Text mining and ontologies in biomedicine: making sense of raw text." Briefings in bioinformatics 6.3 (2005): 239-251.
7. Webber, William. "Evaluating the Effectiveness of Keyword Search." IEEE Data Eng. Bull. 33.1 (2010): 54-59.
8. 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.
9. Santos, José Carlos Almeida, and Manuel Fonseca de Sam Bento Ribeiro. "Improving search engine Query Expansion techniques with ILP."
10. Gregorowicz, Andrew, and Mark A. Kramer. "Mining a large-scale term-concept network from wikipedia." MITRE Corporation 202 (2006).
11. Egozi, Ofer, Evgeniy Gabrilovich, and Shaul Markovitch. "Concept-Based Feature Generation and Selection for Information Retrieval." AAAI. 2008.
12. Tsuruoka, Yoshimasa, Jun'ichi Tsujii, and Sophia Ananiadou. "FACTA: a text search engine for finding associated biomedical concepts." Bioinformatics 24.21 (2008): 2559-2560.
13. 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): e309.
14. 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.
15. Perez-Iratxeta, Carolina, et al. "Update on XplorMed: a web server for exploring scientific literature." Nucleic Acids Research 31.13 (2003): 3866-3868.
16. 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.
17. Abiteboul, Manolescu et al ,”Web Data Management”, Cambridge University Press,2011.
18. Anon, (2014). [online] DL Handbook, Cambridge University Press Available at: http://books.cambridge.org/0521781760.htm
19. Kiryakov, Atanas, et al. "Semantic annotation, indexing, and retrieval." Web Semantics: Science, Services and Agents on the World Wide Web 2.1 (2004): 49-79.

8. ANNEX

[1] The formal terms of the Prader-Willi ontology and the corresponding textual expressions

Formal Terms of the Prader-Willi ontology	Textual expressions
Affabulation	mythomania
AC15P	anomaly of chromosome 15 from paternal origin
Anomaly_Brain_Development_PW	anomaly of brain development, troubles of brain development, brain development disorder
Genetic_Anomaly_PW	genetic anomaly, genetic defect, genetic disorder,genotype
Hormonale_Deficiency_PW	hormonal deficit, hormonal deficiency, pituitary hormonal deficiency, pituitary hormonal deficit, hypothalamic dysfunction,endocrine dysfunction
Anomaly_PW	defect related to Prader-Willi syndrome, anomaly related to Prader-Willi syndrome
Infant_Anorexia	sucking deficit, suckling deficit, failure to thrive, organic failure to thrive of babies, organic failure to thrive of infants, poor sucking, infant anorexia, feeding problem, starvation
Boulimia	boulimia ,hyperphagia ,overeating, over-eating,eating disorder,feeding disorder,food intake
CRH	corticotropin releasing hormone , CRH, CRF factor
ACTH	corticotropin hormone, ACTH

Craving_for_Food	craving for food
Def_empreinte_pat	paternal imprinting deficit, imprinting defect, paternal imprinting mutation, paternal imprinting center mutation,genomic imprinting,imprinting disorder
Satiety_Deficit	deficit of satiety, abnormal satiety, deficit of satiatation
Deficit_of_Social_Skills	deficit of social skills, deficit of social abilities, deficit of social cues, deficit of understanding social codes
Executive_Dysfuntions	executive dysfunctions
Cognitive_Deficit_PW	intellectual disability, cognitive deficit, cognitive troubles, cognitive dysfunctions,cognitive disability,mental retardation, mental deficiency
Deletion_15_pat	chromosome 15 deletion from paternal origin,deletion of chromosome 15Q11
Deletion_SNORD_116	SNORD 116 deletion, SNORD116 gene deletion, cluster SNOD116 deletion, long non coding SNORD116 deletion
Excessive_Day_Sleepiness	excessive daytime sleepiness, sleepiness
FSH	folliculin stimulating hormone,FSH
Food_Addiction	food addiction,hunger
Food_Storage	food storage
Ghrelin	ghrelin,gastric hormone
LHRH	LHRH luteotropin releasing hormone

GnRH	GnRH gonadotropin releasing hormone
Hallucination	delusions, hallucinations
Hormone	hormone
GH	growth hormone, somatotropin hormone, GH
HH	pituitary hormones, hypothalamic hormones, hypothalamic-pituitary hormones
Hyperphagia	hyperphagia , overeating, boulimia
IGF-1	somatomedin, insuline like growth factor 1,IGF-1
IGFBP-3	IGF-1binding protein 3, IGFBP-3
Impulsiveness	impulsive behaviour , impulsiveness
Insulin	insulin
Emotional_Lability	emotional lability, difficulty to control emotion
Hip_Displasia	hip dysplasia
MAGEL2	MAGEL2 gene, MAGEL2
Obesity	obesity, overweight, weight excess, high body fat
Obsession_with_Food	obession with food, permanent thinking for food
Oxytocin	oxytocin, parvotocin, mesotocin, carbetocin
Ophtalmologic_Problems	eyes problems , ophtalmological problems, ocular problems, corneal abnormalities, glaucoma, cataract, corneal nebulae,visual abnormalities,oculo-visual abnormalities, depressed visual acuity, ocular hypopigmentation,strabismus

Prolactin	Prolactin
Psychosis	psychosis, psychotic behaviour/behavior, psychotic signs, obsession
Rigidity_of_Mind	rigidity of mind, mind rigidity
Scoliosis	scoliosis, kyphosis
Skin_Picking	scratching, skin-picking lesions
Hoarding	tendency for hoarding, propensity to hoard propensity to store
Symptom_PW	symptoms, signs, features, criteria, phenotype
TSH	thyrotropin hormone, TSH
TRH	thyrotropin releasing hormone, TRH
Translocation_15_pat	chromosomal translocation of paternal chromosom 15
Compulsive_Behaviour	compulsive behaviour , compulsive behavior, obsessive compulsive disorder, obsessive-compulsive or repetitive behaviors
Troubles_PW	troubles, abnormalities, anomalies, trouble,anomaly,abnormality
Learning_Problems	learning disabilities , learning problems
Sleep_Disorder	sleep disorders , sleep abnormalities
Hyponatremia	hyponatremia, hemodilution, fluid and electrolytes disorders

Gastric_Motility_Problems	gastric motility problems, gastric motility troubles, gastro-intestinal motility problems,vomiting
Apneas	respiratory sleep disorders , apneas , obstructive apnea, central apnea , apnea/ hypopnea index AHI , sleep apnea obstructive syndrome,apnea, dyspnea, respiratory distress
Behaviour_Troubles_PW	behaviour troubles , behavior troubles, behavior disorders, behaviour disorders, abnormal behavior, abnormal behaviour
Eating_Disorder_PW	feeding problems, eating disorders, food intake disorders, appetite regulation problems
Psychiatric_Troubles_PW	psychiatric problems, psychiatric troubles, psychiatric features , psychiatric disorders,obsessive-compulsive disorder
Upd	uniparental disomy, maternal disomy , maternal isodisomy , maternal heterodisomy
Neonatal_Hypotonia	neonatal hypotonia,infantile hypotonia
Mental_Disorder	mental disorder, mental disorders
Idiopathic_Scoliosis	Idiopathic scoliosis
Gene	Gene
IGFBP-7 Gene	Insulin-like growth factor-binding protein 7 gene, IGFBP-7 gene,IGFBP7 gene

MKRN3 Gene	Insulin-like growth factor-binding protein 7 gene, IGFBP-7 gene,IGFBP7 gene
Protein	Protein
IGFBP-7	Insulin-like growth factor-binding protein 7, IGFBP-7,IGFBP7
MKRN3	MKRN3, makorin ring finger protein 3
Necdin	necdin
SNORD115	SNORD115
SNORD109A	SNORD109A
SNORD109B	SNORD109B
IPW	IPW
Leptin	Leptin
Cortisol	Cortisol
Thyroid_Hormones	Thyroid Hormones
Thyroxine	Thyroxine
Triiodothyronine	Triiodothyronine
Animal_Models	animal models
KO mice	KO mice, KO mouse, Knockout mouse, Knockout mice
WT mice	WT mice, WT mouse,Wild type mouse, Wild type mice

Treatments_PW	treatment,therapy
Modafinil_Treatment	modafinil treatment,modafinil therapy
Fluoxetine_Treatment	fluoxetine treatment, fluoxetine therapy
GH_Treatment	GH treatment, GH therapy
Brain_Effects_PW	brain effects
Hypothalamus_Arcuate_Nucleus	hypothalamus arcuate nucleus, arcuate nucleus, infundibular nucleus
Dopaminergic_Circuits	dopaminergic circuits, dopaminergic circuitries
Dopamine	dopamine releases
VTA	VTA, Ventral tegmental area
BDNF	BDNF, Brain-derived neurotrophic factor
Associated_With	is associated to
Stimulates	stimulates
Caused_By	is caused by

[2] SPIN query for ProteinSet_PW

```
CONSTRUCT {
             ?instance a pw:ProteinSet_PW .}
WHERE {{
            ?subject (rdfs:subClassOf)* pw:Protein .
            ?instance a ?subject .}
UNION{
            ?instance (rdfs:subClassOf)* pw:Protein .
            ?instance a owl:Class .} .}
```

[3] SPIN query for AnomalySet_PW

```
CONSTRUCT {
            ?instance a pw:AnomalySet_PW .}
WHERE {{
            ?subject (rdfs:subClassOf)* pw:Anomaly_PW .
            ?instance a ?subject .}
 UNION{
            ?instance (rdfs:subClassOf)* pw:Anomaly_PW .
            ?instance a owl:Class .} .}
```

[4] SPIN query for TroubleSet_PW

```
CONSTRUCT {
            ?instance a pw:TroubleSet_PW .}
WHERE {{
            ?subject (rdfs:subClassOf)* pw:Troubles_PW .
            ?instance a ?subject .}
UNION{
            ?instance (rdfs:subClassOf)* pw:Troubles_PW .
            ?instance a owl:Class .} .}
```

[5] SPIN query for SymptomSet_PW

```
CONSTRUCT {
            ?instance a pw:SymptomSet_PW .}
WHERE {{
            ?subject (rdfs:subClassOf)* pw:Symptom_PW .
            ?instance a ?subject .}
UNION{
            ?instance (rdfs:subClassOf)* pw:Symptom_PW .
            ?instance a owl:Class .} .}
```

[6] SPIN query for GeneSet_PW

```
CONSTRUCT {
          ?instance a pw:GeneSet_PW .}
WHERE {{
         ?subject (rdfs:subClassOf)* pw:Gene .
         ?instance a ?subject .}
UNION{
         ?instance (rdfs:subClassOf)* pw:Gene .
         ?instance a owl:Class .} .}
```

[7] SPIN query for NeurotransmitterSet_PW

```
CONSTRUCT {
         ?instance a pw:NeurotransmitterSet_PW .}
WHERE { {
         ?subject (rdfs:subClassOf)* pw:Neurotransmitter .
          ?instance a ?subject .}
UNION {
         ?instance (rdfs:subClassOf)* pw:Neurotransmitter .
          ?instance a owl:Class .} .}
```

[8] SPIN query for TreatmentSet_PW

```
CONSTRUCT {
         ?instance a pw:TreatmentSet_PW .}
WHERE {{
         ?subject (rdfs:subClassOf)* pw:Treatment .
         ?instance a ?subject .}
UNION{
         ?instance (rdfs:subClassOf)* pw:Treatment .
         ?instance a owl:Class .
} .}
```

[9] SPIN query for HormoneSet_PW

```
CONSTRUCT {
         ?instance a pw:HormoneSet_PW .}
WHERE {{
         ?subject (rdfs:subClassOf)* pw:Hormone .
         ?instance a ?subject .}
UNION{
         ?instance (rdfs:subClassOf)* pw:Hormone .
         ?instance a owl:Class .} .}
```

[10] SPIN query for BrainAnatomySet_PW

```
CONSTRUCT {
             ?instance a pw:BrainAnatomySet_PW .}
WHERE {{
             ?subject (rdfs:subClassOf)* pw:Brain_Anatomy .
            ?instance a ?subject .}
UNION {
             ?instance (rdfs:subClassOf)* pw:Brain_Anatomy .
            ?instance a owl:Class .} .}
```