Ontology Design on Breast Cancer Cause, Symptoms, Diagnosis, and Treatment

Gregory Brown, Sai Sri Narne, Priyanka Gaikwad, and Lakshmi Korrapati

*Department of Computer Science Electrical Engineering*  
*University of Missouri-Kansas City*Kansas City, USA  
gbkhv@ mail.umkc.edu, sn69k@mail.umkc.edu, pvg9gb@mail.umkc.edu, vkvn3@mail.umkc.edu

**Abstract—Ontologies were developed in Artificial Intelligence to facilitate knowledge sharing and reuse. Since the beginning of the nineties ontologies have become a popular research topic investigated by several Artificial Intelligence research communities, including Knowledge Engineering, natural-language processing and knowledge representation. In recent years, there has been explosive growth in the amount of biomedical publications. This paper that deals with the design of an ontology based on causes of breast cancer, Symptoms, identification of Cancer and methodologies used to identify the Cancer in breast cells in order to contribute a knowledge graph in reference to some of the existing ontologies.**

**Keywords — ontology, breast cancer, automated design Spark**

# I. Introduction

Understanding of the cause of disease has improved, as has our capacity to store and process biomedical data. With these advances, the use of complex datasets and applications for processing such datasets present new opportunities to improve our understanding of causes, symptoms, diagnosis, staging, and treatment of disease. We are now able improve the quality of health care decision-making by allowing incorporation of the most recent research results in the decision-making process. Diagnosis and treatment can be guided by development of machine-learning techniques for analysis of the ever-increasing volume of published biomedical research, including unstructured data. Development of an ontological approach to facilitate analysis of this literature would allow more rapid and efficient incorporation of this vast source of data in critical decision-making for patients by physicians and would promote improvements in health care. Specifically, the area of breast cancer causes, symptoms, diagnosis, and treatment, which are the subjects of significant research, would benefit from an ontological approach to facilitate analysis and processing of recently published literature, allowing this literature to be more readily utilized by clinicians to improve patient-oriented outcomes. The focus of this paper is the construction of an ontology using systematic technique for analyzing this literature.

# II. Related Works

Many works have been proposed to create ontologies and their application from a medical corpus. Oshani et al.[4] proposed a prototype system which uses semantic technology approach,which is based on automatic parsing of data from AJCC staging manuals. They have included additional biomarkers which are present in revised 8th edition of AJCC Cancer Staging Manual. The mapping of breast cancer terms to community accepted terms from the National Cancer Institute thesaurus(NCIt) is included as part of ontology creation.

In this work they have considered staging criteria from AJCC 7th edition and AJCC 8th edition to construct Inference agent. The map files which are extracted corresponds to staging criteria specified in AJCC staging manual. For calculation of Inference agent it has not considered the narrower stages specified in AJCC staging manual.

# III. Proposed Work

There are following areas which are focused in our Breast Cancer ontology work.

* Causes
* Symptoms
* Diagnosis
* Treatment

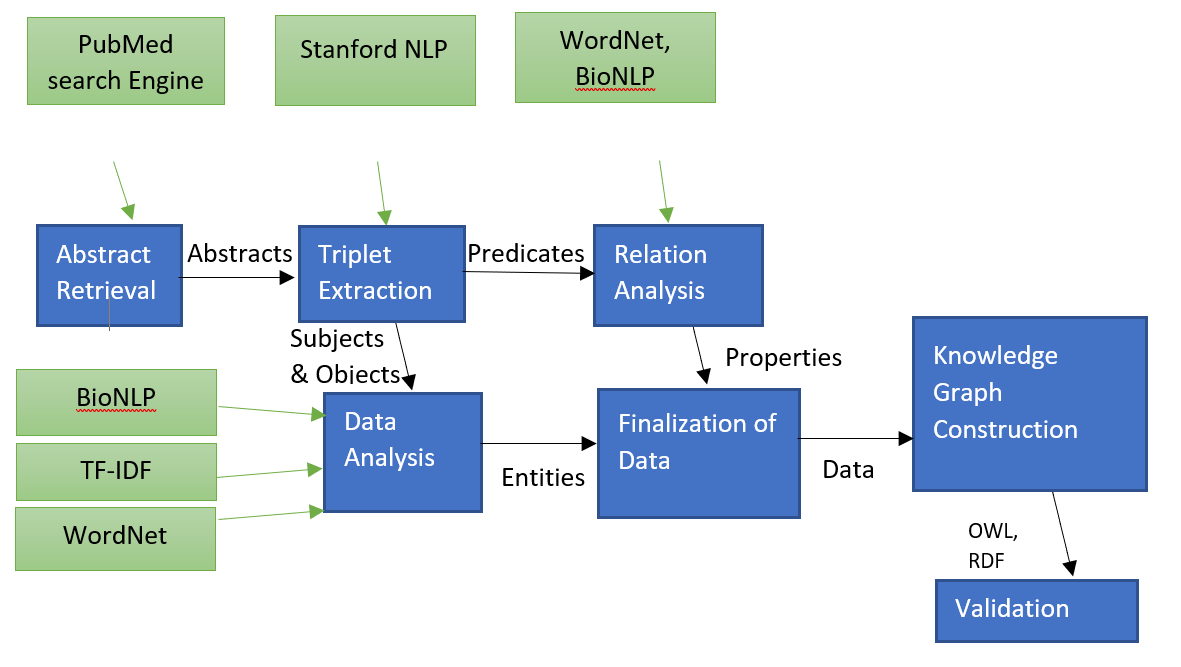


Figure 1: Step-by-step procedure with tools use

The process of generation of Ontology has following steps which are executed sequentially. The flow of these techniques can be seen in figure 1:

1. Information Extraction:

1. Retrieval of abstracts:

There is lot of research work done on breast Cancer by biomedical Scholars and how to integrating results from biolabs to digitalized form, where each of their is of their work is published in the form of some research paper. Here considering such text from these publications processed data for ontology creation. We use PubMed search engine [8] in the Scala code to retrieve PubMedID accordingly to obtain data set of abstracts. NLP actions are performed on this data.

b. Natural Language Processing (NLP):

This is used to process and analyze large amounts of natural language unstructured data. Following processes are conducted as a part of NLP.

* 1. Tokenization is the process of identifying individual words, numbers, and other constructs.
  2. POS Tagging assigns each word in a sentence its respective part of speech such as noun, verb or adjective.
  3. Named Entity Recognition identifies entities such as persons, locations, and times within a text corpus.

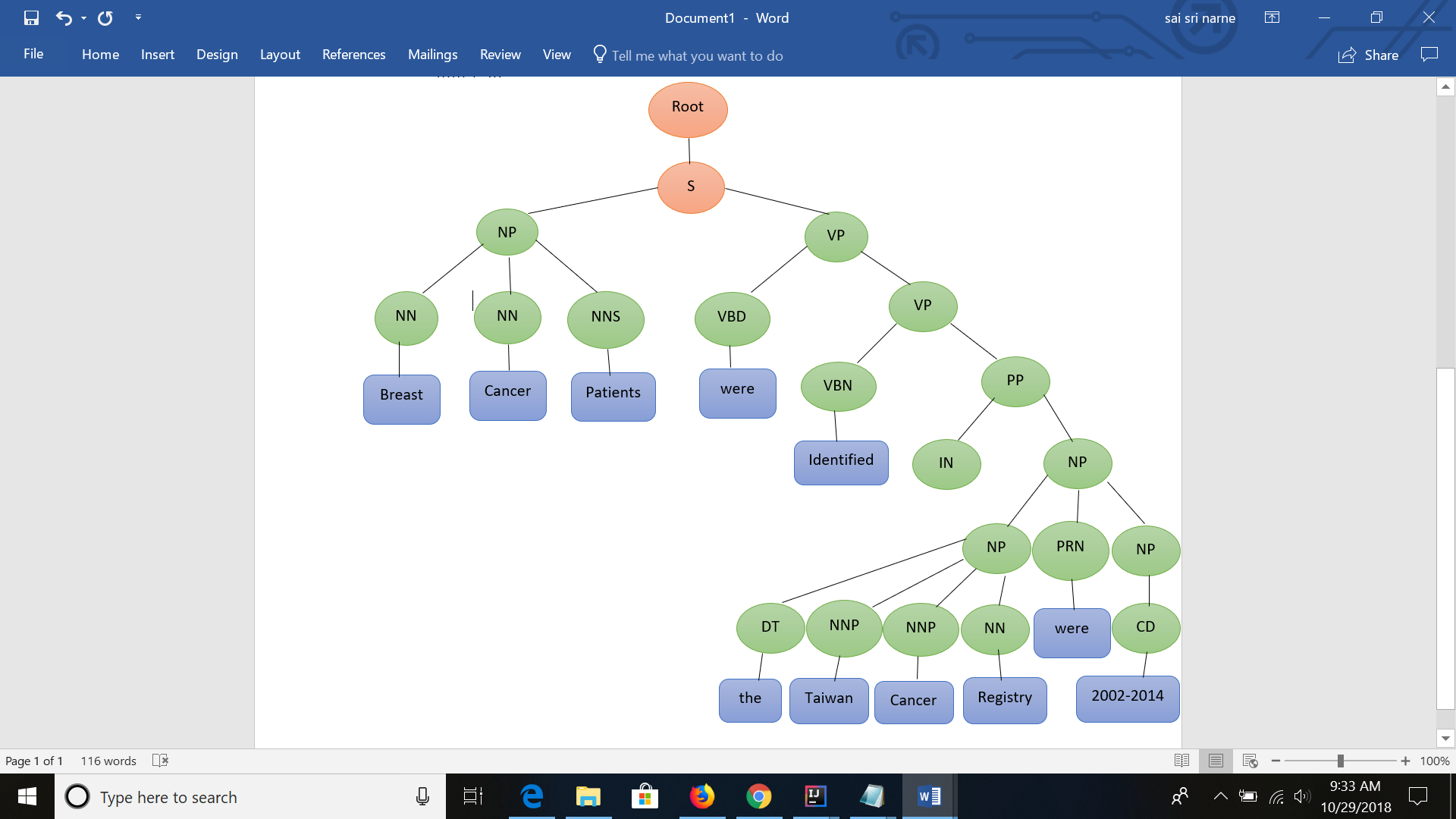
Example sentence: Breast cancer patients were identified from the Taiwan Cancer Registry (2002-2014). Following is the dependencies tree for this sentence. We use NLP for triplet extraction via the Stanford NLP simple document openie function. This triplet extraction is the second step in our ontology construction following retrieval of abstracts. After extracting triplets for each of these sentences we split the triplets into subjects and objects, and predicates.

Figure 2: CoreNLP Basic Dependencies

2. WordNet:

RiTa WordNet provides methods for returning synonyms. We use scripts to find synonyms for words in our subjects and objects and create classes based on commonly occurring synonyms. This same process can be used to simplify our collection of synonyms, though there is no analog of classes and subclasses as data properties, so we must be careful not to lose too much information.

3. BioNLP

BioNLP can be used to find medical words relevant to documents in our corpus based on the PubMed IDs. Medical words belong to one of 5 categories—Chemical, Disease, Gene, Species, or Mutation—each of which is the basis for a class in our final ontology. For our predicates we use BioNLP categorization to determine appropriate domains and ranges for object properties.

After finding classes and subclasses based on our subjects and objects, and properties based on our predicates we filter to remove duplicates and output these classes, subclasses, and properties along with the triplets and individuals to text files which will be used in ontology construction.

4. Ontology Construction:

The Semantic Web organization provides a library which includes functions for the construction of an ontology. Our ontology construction process incorporates the data from the text files containing classes, subclasses, properties, individuals and triplets and constructs an OWL xml ontology using the functions provided by the Semantic Web organization.

5. Information Retrieval:

We have used information retrieval techniques to process the data and provide structure to our ontology. Specifically, we group individuals in our final ontologies into groups based on comparison with top TF-IDF terms. We got the Rita WordNet Synonyms for these top TF-IDF words and if the entities contained either a TF-IDF term or a synonym then that entity was grouped into that class. We could have used word2Vec to find synonyms rather than Rita WordNet but the results we got from Word2Vec were not promising. Interestingly, these information retrieval techniques to do not These techniques do not refer to any outside sources or known sentence structures, instead analyzing the mathematically.

1. TF-IDF combines term frequency and inverse document frequency of terms in the corpus to determine which terms are relevant.
2. Word2Vec creates vectors, or dynamic arrays, by creating a co-occurrence matrix and performing single value decomposition. The resulting vectors are analyzed using cosine similarity to determine which terms in the corpus are synonymous

**Algorithms and Pseudocode:**

Much of the code for the retrieval of abstracts is done via API calls or use of NLP libraries. In particular for analysis of parts of speech and lemmatization we use the Stanford CoreNLP library and for retrieval of synonyms (prior to word2Vec) we use Rita WordNet library functions. Since Stanford CoreNLP is a Java library, we have used auxiliary scripts in Java in our Scala code.

|  |
| --- |
| **Algorithm1** |
| **Ontology\_Construction**()  {  Abstracts = **RetrieveAbstracts**(PubMedIDs)  CoreNLPResults[] = **StanfordCoreNLP**(Abstracts)  Triplets = CoreNLPResults[1]  BioMedical\_Objects = **BioNLP**(Triplets.objects)  BioMedical\_Subjects = **BioNLP**(Triplets.subjects)  Subject\_Synonyms = **WordNet**(Triplets.Subjects)  Object\_Synonyms = **WordnNet**(Triplets.objects)  Subject\_Considered = combine(BioMedical\_Subjects, Subject\_Synonyms)  Object\_Considered = combine(BioMedical\_Objects, Object\_Synonyms)  Predicates\_Considered=compare(Triplets.predicates)  Trips = combine(Subject\_Considered,  Object\_Considered,  Predicates\_Considered)  Triplets\_Considered = compare(Trips, existing\_ontology\_triplets)  OWL = generate\_OWL(Triplets\_Considered)  Visualaize\_Ontology(OWL)  } |

# IV. Software &System Architecture

**Software:**

In this paper we carried our work by Intellij IDEA Community Edition 2018.2.2 environment using Scala SBT version1.0 and Spark 2.12.0 framework. Scala largely provides support for functional programming and static types making lives easier. In order to efficiently perform the operation over large corpuses we used spark framework to divide the entire corpus into small chunks of data and process it in parallel across partitions of our processing power, emulating the Hadoop distributed file system (HDFS). We set the memory and number of partitions to be used in order to utilize more of our machines resources than might otherwise be available. Notably, the pictured Spark initialization is at the head of our script which gets medical word statistics via API call.

**System Architecture:**

In-order to achieve our goal there are many preprocessing steps to be followed as in figure-2. The following are the tools used:

PubMed search engine- In the first step of figure, we used PubMedID’s to retrieve abstract text. Since, PubMed has over 28 million citations of biomedical information from MEDLINE and also a lot of journals, we used this tool to extract data. It is a free resource provides access to United States National Library of Medicine Database [8]. It is also maintained by National Center of Biotechnology Information(NCBI). This search engine makes our process to extract all the literature that contains medical information.

Stanford CoreNLP- As a part of Natural language processing, text annotation, named entity recognition, parts of speech are to be performed on the extracted text or also called to be the linguistic analysis of data, we made use of Stanford CoreNLP [9]. Stanford NLP group has come up with Stanford CoreNLP, a natural language software as their goal to make it very easy for performing linguistic analysis tools to a piece of text using pipeline. The results we got are quite interesting and also accurate mentioned in the table 1

National Center of Biotechnology Information (NCBI)- In general the medical words comprise of various categories like species, gene, chemical etc. NCBI [10], which is a part of United states of National Library of Medicine allows us to search for proteins, minerals, gene, species, chemicals etc. We made sure to find category of each word.

National Center of Biomedical Ontology (NCBO)- There might be some existing list of ontologies on specific topic, NCBO [11] helped us to find the list of ontologies providing BioPortal Rest API call.

WordNet- In order to analysis more about the text data, we used WordNet. WordNet is a large lexical database of English sometimes called as a dictionary that combines nouns, verbs, adverbs, adjectives into a cognitive set of Synonyms (Synsets). We analyzed these results and compared this with Word2Vec and term frequency Id of words, shown in table 2.

Semantic Web- It is an idea of world web inventor, a mesh of data that can be processed by machines easily. Semantic web [12] makes machines work like intelligent humans. We developed W3C web ontology (OWL) to represent rich and complex knowledge about breast cancer cause, treatment, diagnosis using the extracted data.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Words | Unique Words | Nouns | Verbs | Other POS | WordNet Words | Unique Wordnet Words | Medical Terms | Unique Medical Terms |
| 17267 | 3959 | 3960 | 1421 | 4289 | 8559 | 1868 | 679 | 204 |

Table 1: data statistics

# V. Implementation and Evaluation

We use scripts throughout the analysis of the corpus to automate the process. Currently the steps described in the proposed work section of this paper are executed by collections of scripts, however ultimately, we intend to streamline this. The scripts used are composed in a combination of Java and Scala with Spark. Thus far our corpus consists of 99 abstracts retrieved from PubMed collected automatically by providing query terms, specifically breast cancer causes, symptoms, diagnosis, and treatment. However, we have manually vetted these abstracts to some extent. First, we analyzed the corpus using NLP to get data statistics shown in table 1.

Code used to process and analyze corpus is stored at: <https://github.com/toadSTL/CS5560BreastCancerOntology>

These data statistics are obtained via spark map-reduce functions which map ‘<word>’ to ‘(<word>,1)’ for words in the corpus and then reduce by key. For part-of-Speech Statistics we use Stanford Core NLP pipeline to annotate the data, and then do map-reduce, and similar techniques are used for WordNet word analysis and Medical term analysis.

These data statistics are not used for ontology construction, but are helpful in understanding the data which we continue to process. The major steps of our ontology construction are laid out in section III of this paper. In the remainder of this section we will follow an example of what this process does to a single sentence to show the process. The following the sentence we will use as the example:

“Sentinel node biopsy is now also widely used in patients with pathologically positive nodes who receive neoadjuvant chemotherapy.”

The first step is to do triplet extraction on the sentences in the corpus. Our script for this reads a fresh instance of the corpus into a Spark context so that we can first perform triplet extraction. For the example sentence the following triplets are returned:

[(Sentinel node biopsy,is,now widely used,1.0), (Sentinel node biopsy,is now used in,patients,1.0), (Sentinel node biopsy,is now widely used in,patients,1.0),

...

, (Sentinel node biopsy,is,used,1.0), (Sentinel node biopsy,is now also widely used in,patients,1.0)]

In total 17 triplets are initially returned. Of these triplets we filter so that we at most three triplets: the shortest overall triplet; the longest overall triplet; and the triplet with the longest subject and objects as well as the shortest predicate. This filtering leaves us with:

[(Sentinel node biopsy,is,used,1.0), (Sentinel node biopsy,is,now also widely used,1.0), (Sentinel node biopsy,is now also widely used in,patients,1.0)]

Going forward in explanation of this example we will consider the third of these triplets. Thus. the next step performed by our scripts is to use RiTa WordNet and BioNLP to categorize the subject and object into classes. BioNLP determines ‘patients’ to be a species. ‘Sentinel node biopsy’ is found to belong to the class ‘Biopsy’ by the TF-IDF information retrieval described below.

At this stage we perform Information Extraction techniques to help us improve our class structure. We have

performed the Information Extraction techniques TF-IDF on the subjects and objects our corpus and the top results provide base classes into which we group our individuals. When we do TF-IDF. The result of TF-IDF is a matrix corresponding to the terms from the corpus. The below tables contain the top TF-IDF terms and a selection of the individuals which fall under the classes that we use those terms to generate:

|  |  |
| --- | --- |
| **TF-IDF Classes** | **Individuals** |
| **Molecular** | Molecular Evolution Assay, directMolecularLinksBetween  direct molecular links, |
| **altitude** | highAltitudeMestizoPopulation, altitudePopulation |
| **rates** | demonstrates |
| **implications** | implicationsForTreatment, helpfulImplications, |
| **population** | manyDifferentPopulation,  variousPopulationsIncludingThoseVerySimilarToVietnamese |
| **risk** | riskAllele  breastCancerRisk  increasedRiskOfMortality  riskForPatientsDiagnosed  increasedRisk |
| **characteristics** | associatedWithCharacteristics |
| **absence** | absenceInEjaculate  absenceOfSpermatozoaInEjaculateDueToSpermatogenesis |

Table 2: TF-IDF Classes and Individuals found via 'Causes'

|  |  |
| --- | --- |
| **TF-IDF Classes** | **Individuals found ‘Treatment’** |
| **Treatment Plans** | EgKI-1 treatment, Adjuvant breast radiotherapy , Co-treatment, free-radical scavenging, Association of chemotherapy, ERI 66, Extended endocrine therapy, initiation of chemotherapy term, ERI degradation, ERI transcription cycle |
| **Anticancer Activities** | Oncosphere of tapeworm Echinococcus, anti-cancer therapeutic, activity in in-vitro models, Intraclass correlation coefficient |
| **In Vitro Models** | Time-enhancement curves, Activity in in-vitro models, Uniform random distribution,  corresponding curves |
| **Adapter Protein** | Role on IR signaling, IR mRNA expression levels, expression of mitochondrial apoptotic proteins, coactivators, ERI protein levels |
| **Screening Methods** | QPCR screening, Screening Tools |
| **Cell Cycle Activities** | Normal cell growth , cell cycle, cell proliferation , cell migration, cell apoptosis,G2/M phase, Cell growth of MCF-7 cells at higher concentrations |
| **Breast CT perfusion** | Breast CT image, 4D breast CT image acquisition, Breast tissue perfusion |
| **BioMarkers** | potential biomarkers of resistance, HER2-positive BCs |
| **Treatment Outcomes** | Cosmetic outcomes, late recurrence risk, Memory decline, Upregulated expression |

Table 3: TF-IDF Classes and Individuals found via 'Treatment'

|  |  |
| --- | --- |
| **TF-IDF Classes** | **Individuals** |
| **Biopsy** | tissueBiopsy, sentinelLymphNodeBiopsy, biopsy openBiopsyOfNodularityAttachedToPectoralMuscleFascia, stereotacticWire-guidedBiopsy, rightMastectomyWithSentinelLymphNodeBiopsy |
| **Hazard** | Risk, anticipatedOutcome, lowRiskLesions, subjectiveRisk, treatmentOutcomePrediction, riskFactor, poorerOutcome, associatedWithBreastCancerRisk, RiskProtection |
| **Hereditary** | geneticMutation, geneticChanges, differentGeneticProfile, hereditaryCausesOfChorea, identificationOfDistinctiveGeneticModificationsInDifferentBreastCancerMolecularSubtypes |
| **Protocols** | differentProtocols, protocols, treatmentProtocols |

Table 4: TF-IDF Classes and Individuals found via 'Diagnosis'

As the table shows, some of the top TF-IDF sometimes provide good categories, but not always. Some manual vetting was does of TF-IDF terms. For example, for terms found via ‘Diagnosis’ terms already found as medical words were removed as well as adjectives and other non-noun terms. Furthermore, for terms found via ‘Diagnosis’ top TF-IDF terms which were deemed ‘too-specific’ were not used as classes. These classes help greatly in developing the structures of the ontologies we generated.

Continuing with our example we prepare the subject, predicate, and object from our sentence for inclusion in our ontology. At this stage we remove stopwords and remove duplicates, but this triplet doesn’t contain any stopwords and without loss of generality we can suppose that this is the version of that triplet which was not removed as a duplicate. Preparing it for inclusion in our ontology involves adding the class and subclasses relevant for this triplet to our lists of classes and subclasses. In this case we add ‘Gene’ and ‘Disease’ as classes and add ‘CXCR4’ as a subclass of ‘Gene’ and ‘invasiveBreastCancer’ as a subclass of ‘Disease’. In addition to adding these items to our subclasses we also currently add them as individals. Please not that for invasive breast cancer we have converted the phrase to camel-case, and removed spaces, we did the same for ‘CXCR4’ but this process does not change that entity’s name. We also add the line ‘overexpressedIn, Gene, Disease, Obj’ to our list of object properties as Gene is the domain for this function and disease is the range. This is determined by comparing the triplet with separate filtered lists based on our classes. Finally, we add the triplet to our triplets output document using the following line:

‘CXCR4,overexpressedIn,invasiveBreastCancer,Obj’

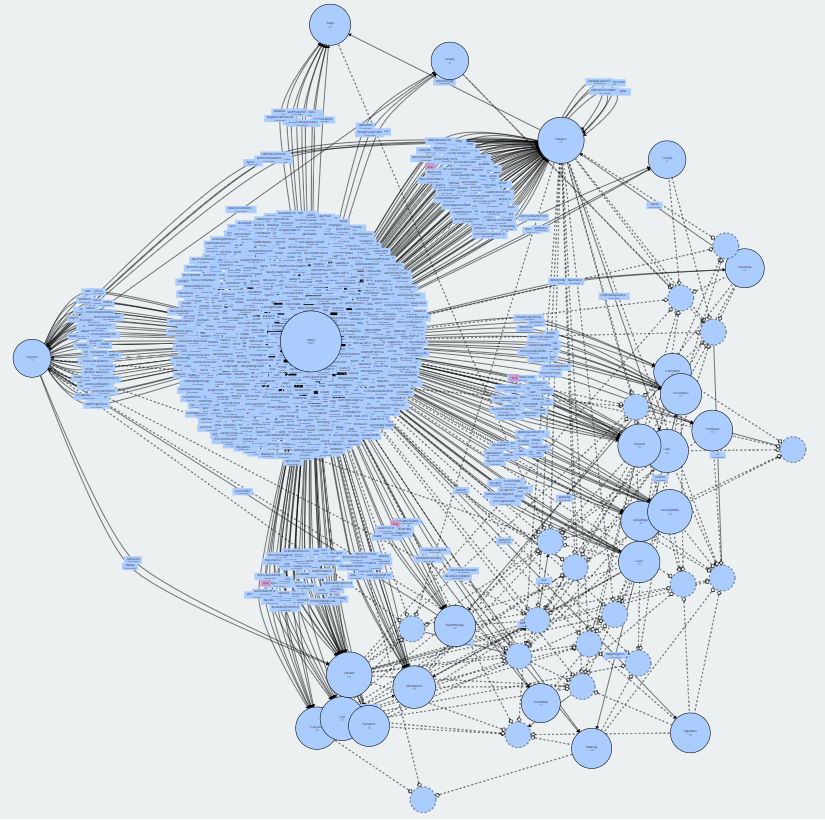
Once we have output all of our data to text files we use the Semantic Web organization’s function for ontology creation to generate an ontology based on the data. Fig. 3 shows a visualization of the Breast Cancer Diagnosis ontology we have constructed and Table 5 shows statistics about that ontology as well as the Breast Cancer Causes Ontology and the Breast Cancer Treatment Ontology. Similarly, Fig. 4 and Fig. 5 show the ontologies we constructed for Breast Cancer Causes and Breast Cancer Treatment, respectively.

Figure 3: Ontology Visualization for Breast Cancer Diagnosis

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Ontology | Classes | Object Properties | Individuals | Nodes | Edges |
| Diagnosis | 269 | 953 | 3370 | 269 | 1506 |
| Causes | 447 | 90 | 34 | 447 | 414 |
| Treatment | 132 | 557 | 2587 | 132 | 819 |

Table 5. Statistics about Breast Cancer ontologies

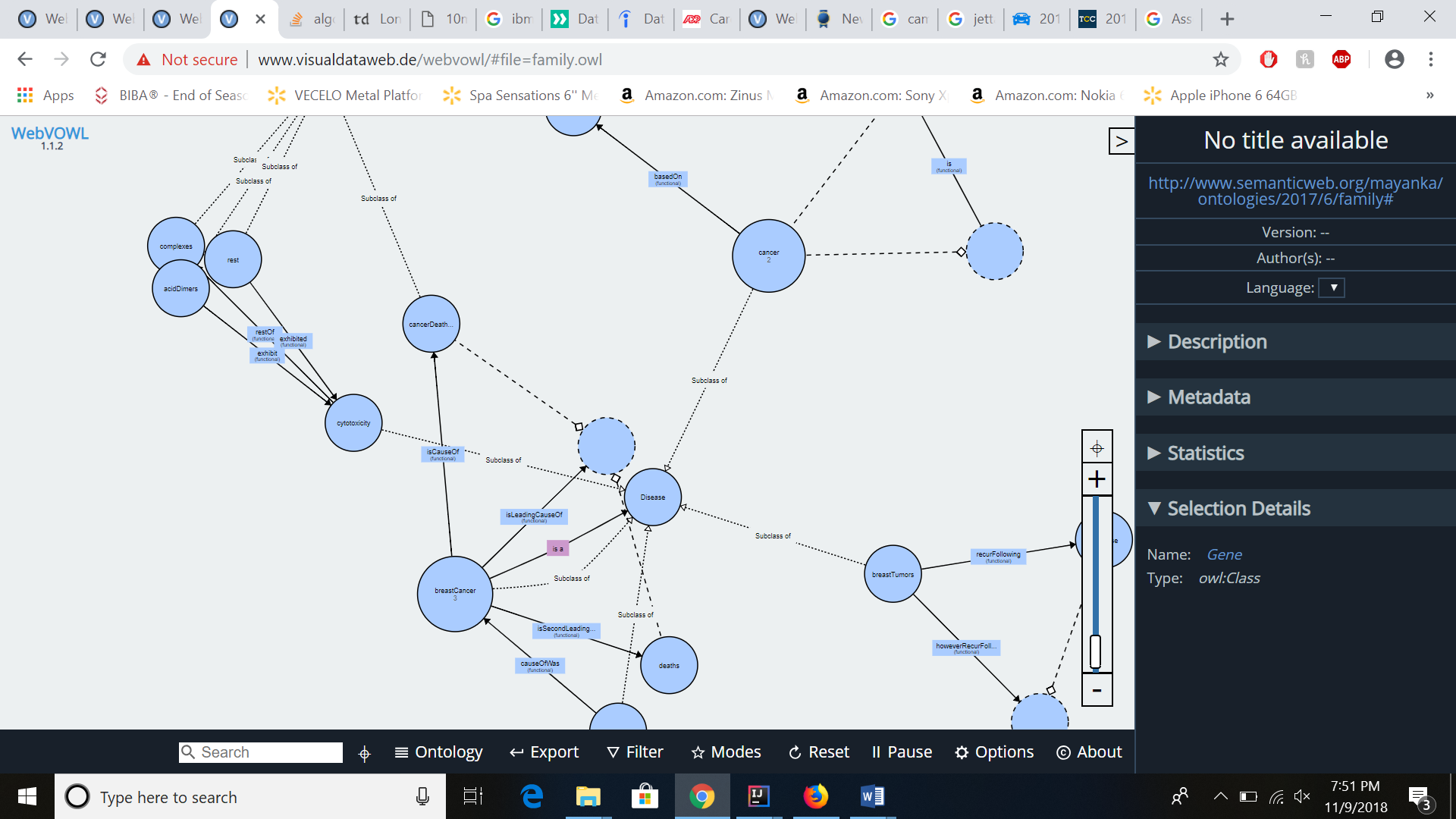
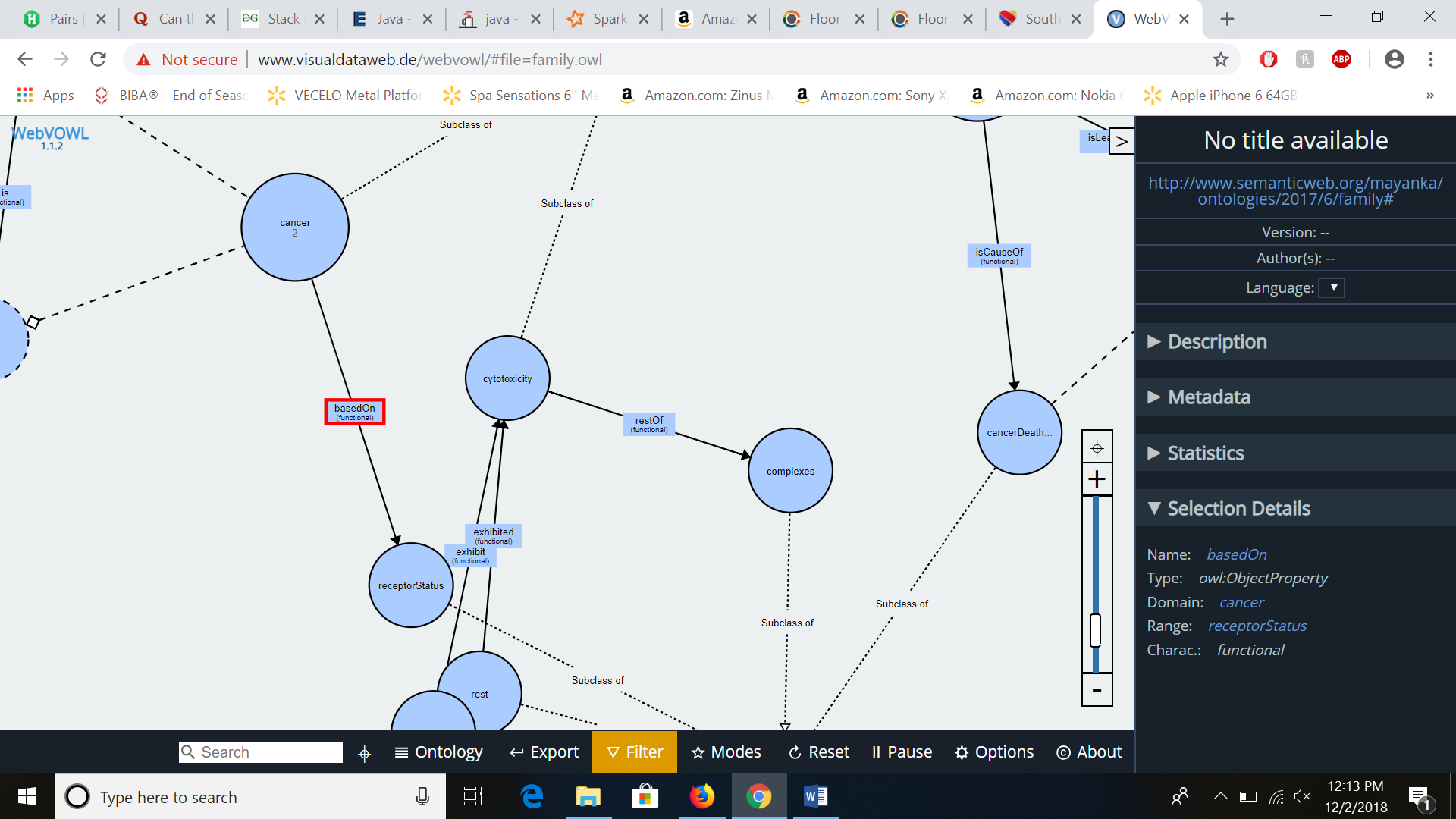


Figure 4. Section of Ontology Visualization for Breast Cancer Causes



*Figure 5. Representation of the properties*

Domain and range of object property lies in between sub class and subclass for example consider “subclass cancer of the class disease and it has the individual tumor has the property basedOn has domain as cancer and range as receptorStatus.

Apart from the object properties, some other properties like transitive, inverse and symmetric which are more significant in OWL by W3C. We figured out that for breast cancer cause data symmetricProperty rule worked for better in reasoning using rule based reasoning. For example, consider we have one sentence “ breastCancer is leadingCause” symmetric to “leadingCause is breastCancer”. Some of the classes in our ontology had partial matched with Ontology for breast cancer treatment. Though the class are not be clearly seen in figure 6 visualization ontology.

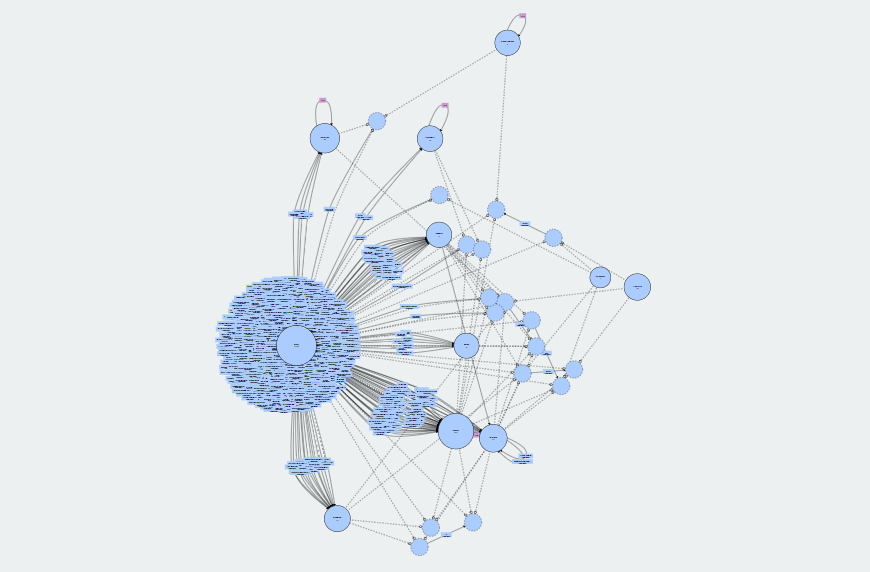
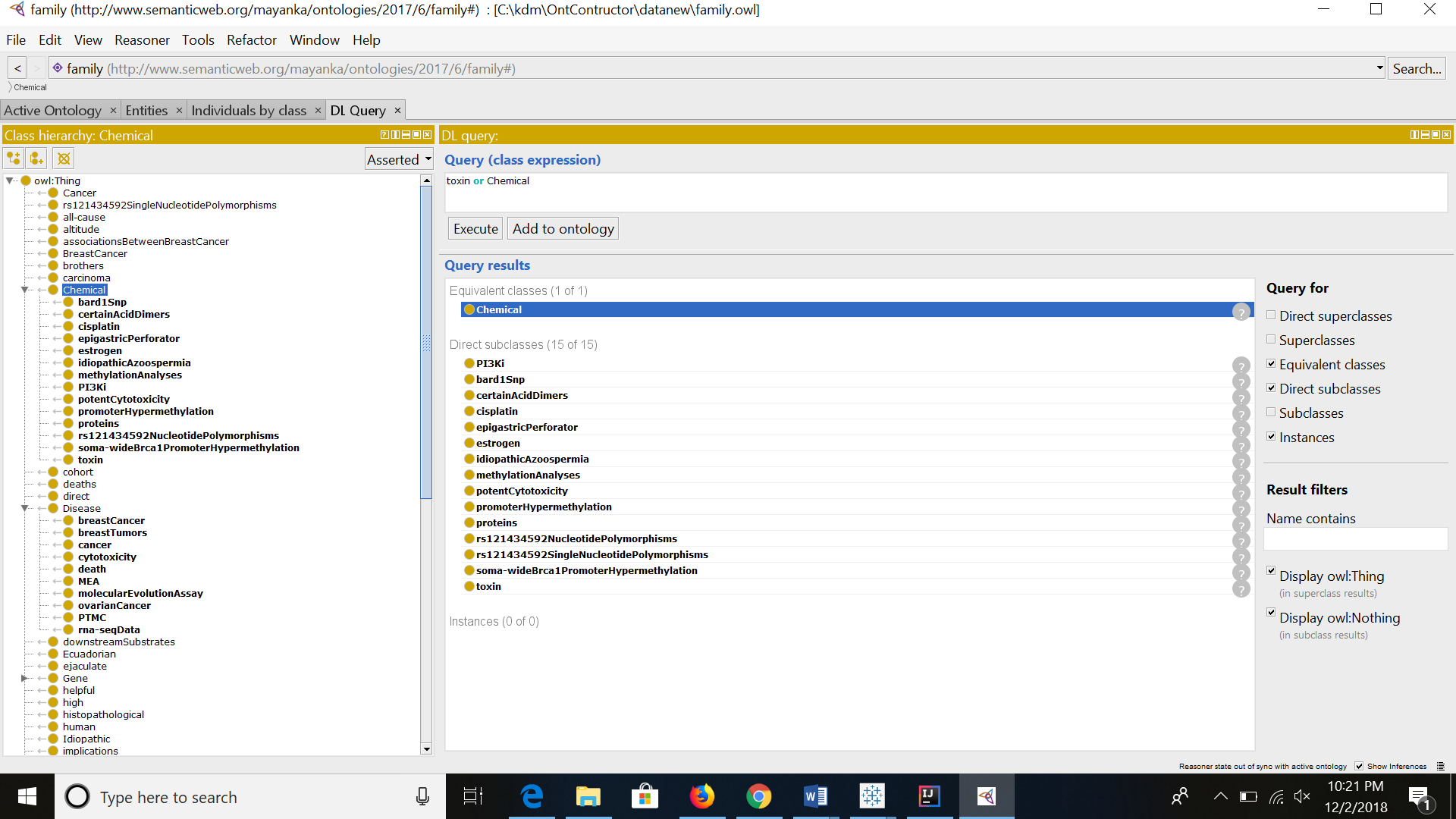


Figure 6: Ontology visualization for Breast Cancer Treatment



*Figure7 DL Query results Analysis*

We also performed some query processing on the ontology by loading OWL file in Protégé tool For example, the query “toxin or Chemical” where toxin is the subclass of Chemical and Chemical is class. The above figure shows the output of the query like it represents all the direct subclasses and some instances.

Additionally, we included images word clouds of top TF-IDF terms for these se different topic in figures 6 through 8. Figure 6 shows a word cloud TF-IDF terms which were used as classes with size corresponding to the number for individuals belonging to class for that TF-IDF term. As can be seen through this visualization for the most part these TF-IDF classes are, or can be understood as being, medically 

Figure . Word Cloud for TF-IDF terms based on 'Diagnosis'

relevant. A few examples are ‘Disease’, ‘Cell’, ‘Care’, ‘Symptom’, and ‘Mechanism’ which are all potentially relevant to Breast Cancer Diagnosis.

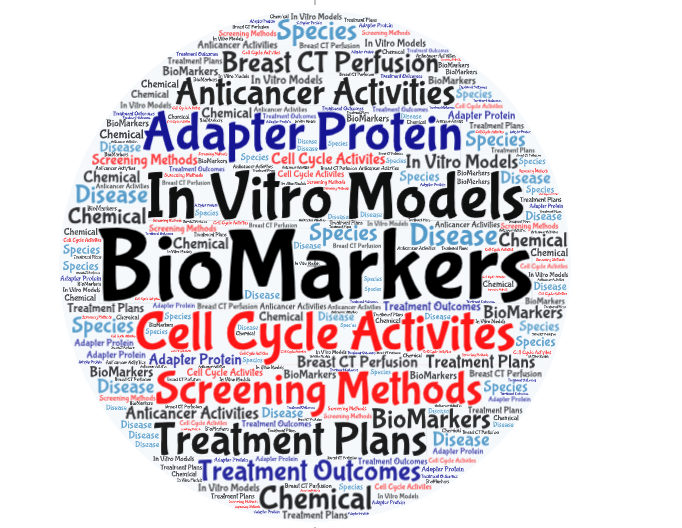


Figure 8: Word Cloud for TF-IDF terms based on 'Treatment

Figure 8 shows a word cloud TF-IDF terms for Breast Cancer Treatment which were used as classes with size corresponding to the number for individuals belonging to class for that TF-IDF term.TFIDF process is used to get these words and wordgenerator is used to create this wordcloud.

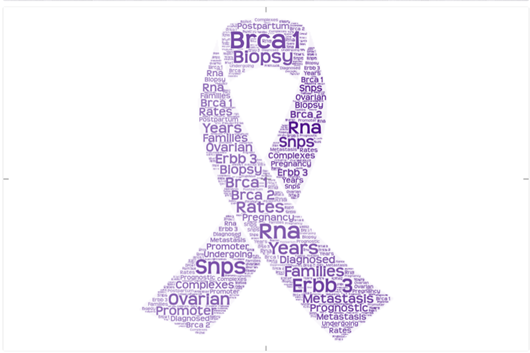


Figure 9: Word Cloud for TF-IDF terms based on 'Symptoms’

Finally, we have included visualizations of ontologies to which we compare the ontologies we have generated. Specifically, only show two visualizations of related ontologies because two of the ontologies which we are comparing are difficult to visualize.

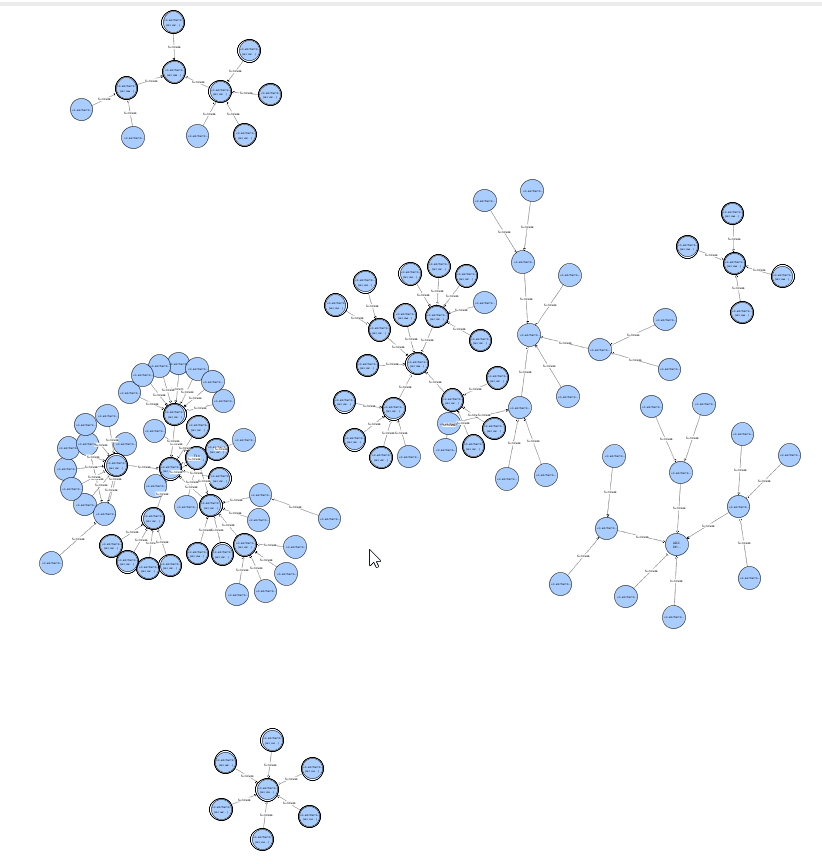


Figure 10: Visualization Breast Cancer Staging Terms Ontology

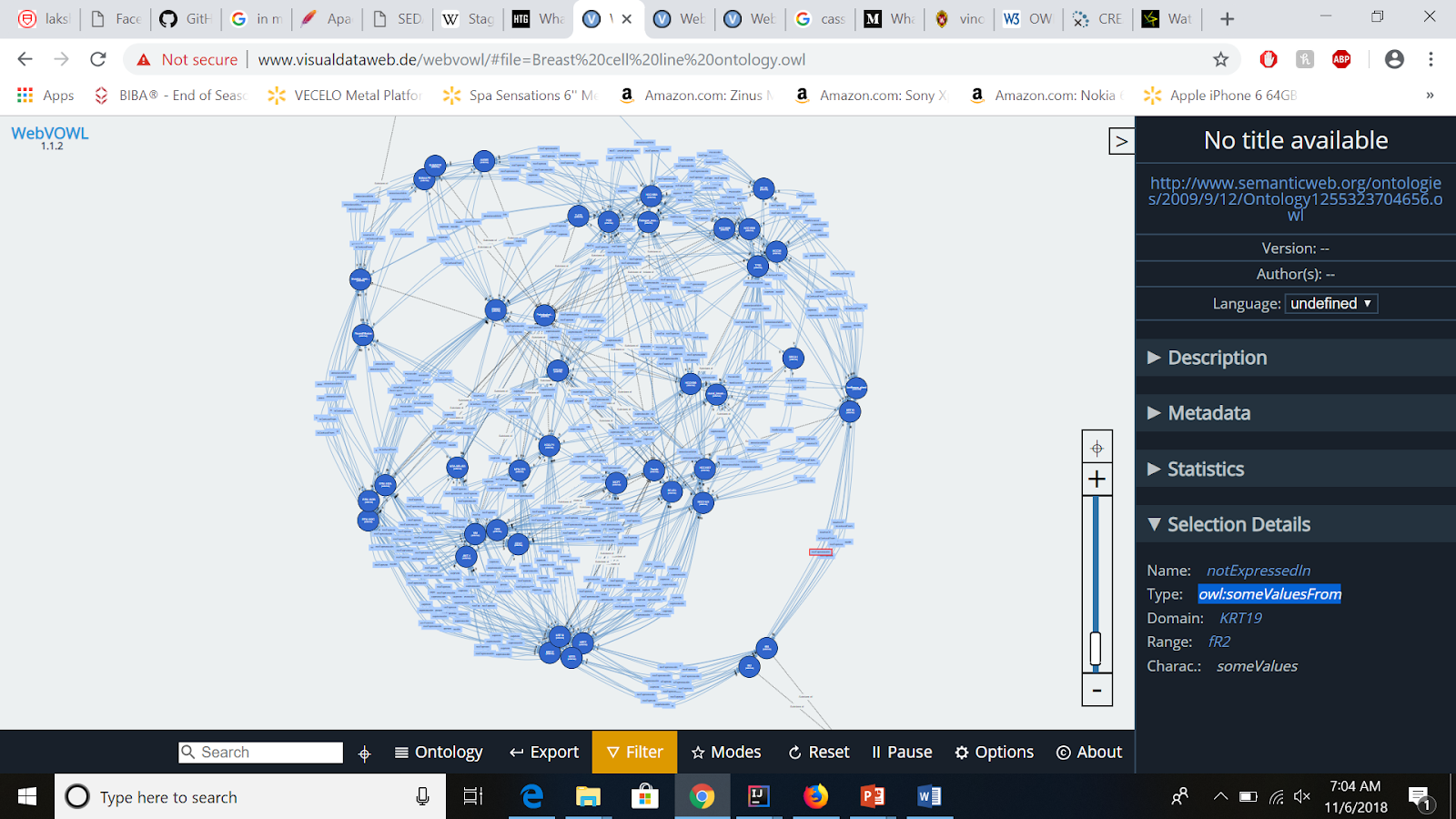


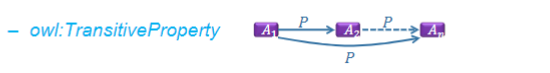
Figure 11: Visualization related ontology

The Visualization related ontology for Breast Cancer Treatment is not included as the ontology does not resemble any of the classes under Breast Cancer Treatment Ontology.

When we compare Breast Cancer Treatment Ontology it specifies treatement outcomes rather the treatment plans.

We also performed properties on Ontology to determine rule based triplets which specifies following properties for triplets.Some of the properties are explained in below examples.

Transitive Property:



Following are the triplets from Breast Cancer Treatment Ontology satisfies the transitive property.

mastectomies,isWith,immediateBreastReconstruction  
immediateBreastReconstruction,isWith,implants

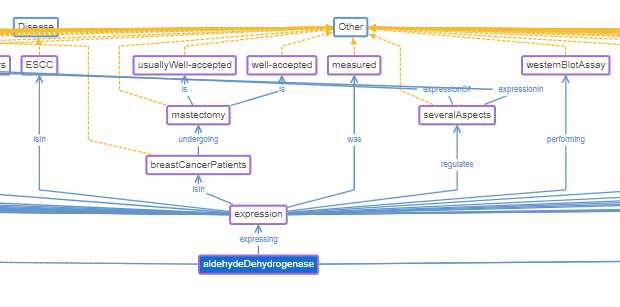
Transtive :  
mastectomies,isWith,implants

breastCancer,is,cancer  
cancer,is,treatable

Transtive :  
breastCancer,is,treatable

hypofractionatedVmat-sib,isIn,patients  
patients,isIn,ebrtGroup

Transtive :  
hypofractionatedVmat-sib,isIn,ebrtGroup



Taking a closer look at a particular example to see the kinds of structures arising consider the sentence from our corpus for Breast Cancer Treatment Ontology.

aldehydeDehydrogenase is individual which is considered for above strurural relation.

Corresponding triplet is as follows:

aldehydeDehydrogenase,expressing,expression

Above figure shows relationship between expression with other entities in our ontology.

One ontology which we have not included a visualization of is the NCIT or National Cancer Institute Thesaurus. Although we are not able to show a visualization of this complete ontology, we can compare the class structure of NCIT with our Breast Cancer Diagnosis ontology. In doing so we find that many of the highest-level classes in the NCIT are equivalent to classes of our Breast cancer Diagnosis Ontology. Classes of the NCIT ontology for which there are particularly good matches include the Activity; Chemotherapy Regimen or Agent Combination; Diagnostic or Prognostic Factor; Disease, Disorder or Finding; Gene; and Organism. These classes from the NCIT ontology are respectively comparable with the following classes from the Breast Cancer Diagnosis ontology we constructed: Activities; Radiotherapy; Factor; Disease; Gene; and Species.

While these classes from our Breast Cancer Diagnosis ontology ultimately do contain some of the same individuals as their corresponding classes in the NCIT ontology, the NCIT ontology has added structure, for example breaking down their Gene class into subclasses by categorizing genes. Furthermore, there are different naming conventions implemented in the NCIT ontology. Nonetheless genes from our Gene class do frequently appear in under some subclass of the NCIT Gene class, for example: our Gene class includes microRNA-22, where the NCIT Gene class contains MIR22 Gene.

Another area of analysis we performed on the ontologies we constructed was to find relations, or object properties, which had certain properties. For example, in our Breast Cancer Diagnosis Ontology we determined that both the ‘is’ and the ‘isIn’ relations were transitive.

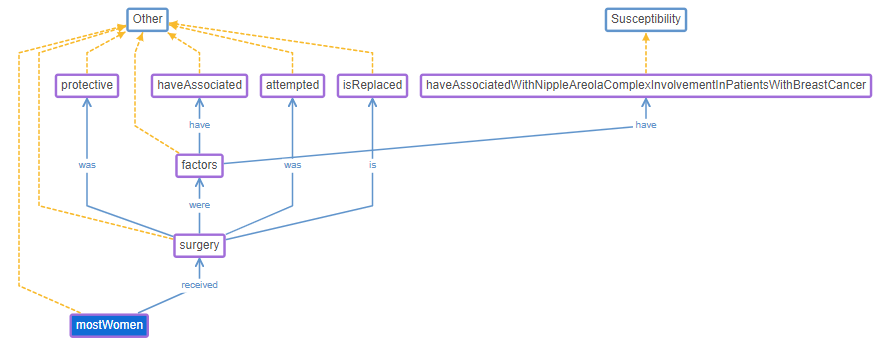
Taking a closer look at a particular example to see the kinds of structures arising consider the sentence from our corpus:

Most women received surgery: 72% BCS; 24% mastectomy.

This sentence is translated into the following triplet through our ontology construction process:

mostWomen, received, surgery

Here mostWomen and surgery are Individuals in our ontology and received is an Object Property. Below is a visualization of the relationships associating the Individual mostWomen with other entities within the ontology.



Ultimately there are still many issues with our Breast Cancer Diagnosis ontology. The most striking of these issues is that the majority of Individuals still fall into the Other class. Another major issue is that there is still no, class-subclass structure. Finally, while some of the entities and relations are appropriate and meaningful still many meaningless, or at least confusing, entities and relationships still exist within the ontology.

# V. Conclusion

In this paper, the major work is focused on automated development of ontologies on breast Cancer using PubMed publications. Ontologies constructed include of breast cancer cause, symptoms, diagnosis and treatment. The first step in our process of constructing these ontologies was to retrieve abstracts from PubMed, filtering out irrelevant and undesirable abstracts. The next step was to do triplet extraction on the texts from these abstracts using Stanford CoreNLP functions, and again filtering was done to reduce the total number of triplets returned. TF-IDF was then performed on the subjects and objects from the accepted triplets. The top TF-IDF terms were then manually filtered and the resulting list was used as classes into which to place individuals during the ontology construction phase, associated with synonyms via Rita WordNet. The final step in pre-processing the data was to retrieve triplets again, this time grouping subjects and objects as individuals into classes based on BioNLP terms and the previously mentioned TF-IDF terms. Additionally, predicates are associated to domains and ranges from the lists of classes and saved as object properties for the ontology construction. The resulting ontologies bare some resemblance to existing ontologies on topics surrounding breast cancer, however, the ontologies we constructed leave much to be desired.

Further research could be done on this topic refining the process of ontology construction to produce better results and further automate the process. Specifically, additional subclass structure would greatly benefit these ontologies. Techniques which could be used to generate such structure or otherwise improve the ontologies we constructed include LDA, Word2Vec, further TF-IDF, stemming and lemmatization, natural language processing of the subjects, objects, and predicates in order to determine more robust set of individuals and object properties, analysis of predicates and noun-clauses as possible candidates for data properties or data, respectively and any combination of the aforementioned techniques. Finding the best possible pipeline for ontology construction, even specific to the domain of Breast Cancer would be time consuming and would not necessarily be generalizable. But this further research can be motivated, by the increasing volume of medical data published, and the need organize that data and make it usable for doctors and medical researchers, by way of clinical decision assistance software drawing on ontologies.

# References

1. <https://ieeexplore.ieee.org/document/8217715>

1. <http://dline.info/fpaper/jdim/v15i5/jdimv15i5_3.pdf>
2. <https://arxiv.org/pdf/1607.08074.pdf>
3. <https://arxiv.org/pdf/1807.07991.pdf>
4. <https://ieeexplore-ieee-org.proxy.library.umkc.edu/document/4562013/>
5. <https://ieeexplore.ieee.org/document/6834613/>
6. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5174013/>

[8] <https://www.ncbi.nlm.nih.gov/pubmed/>

[9] <https://stanfordnlp.github.io/CoreNLP/>

[10] <https://www.ncbi.nlm.nih.gov/>

[11] <https://bioportal.bioontology.org/>

[12] <https://www.w3.org/standards/semanticweb/>

# WORK

* Causes Lakshmi Korrapti
* Symptoms Sai Sri Narne
* Diagnosis Gregory Brown
* Treatment Priyanka Gaikwad