Gregory Brown

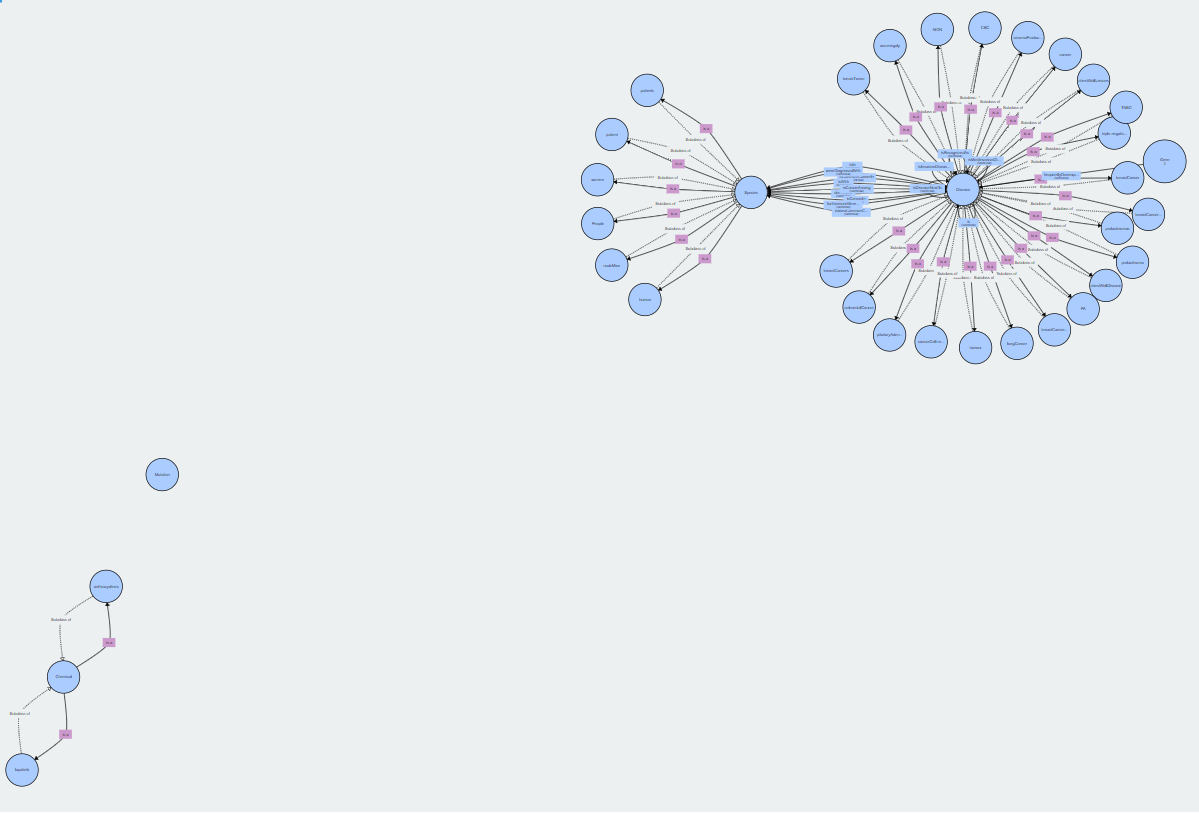
Class ID: 1

[Code](https://github.com/toadSTL/CS5560GregBrownLabs/tree/master/Lab3)

**1) Data statistics from previous ontology:**

Version 1 Statistics (BreastCancer.owl):

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Triplets | Subjects | Objects | Predicates | Classes | Subclasses | Individuals | Object Properties |
| 1507 | 312 | 1660 | 395 | 6 | 1883 | 1850 | 395 |

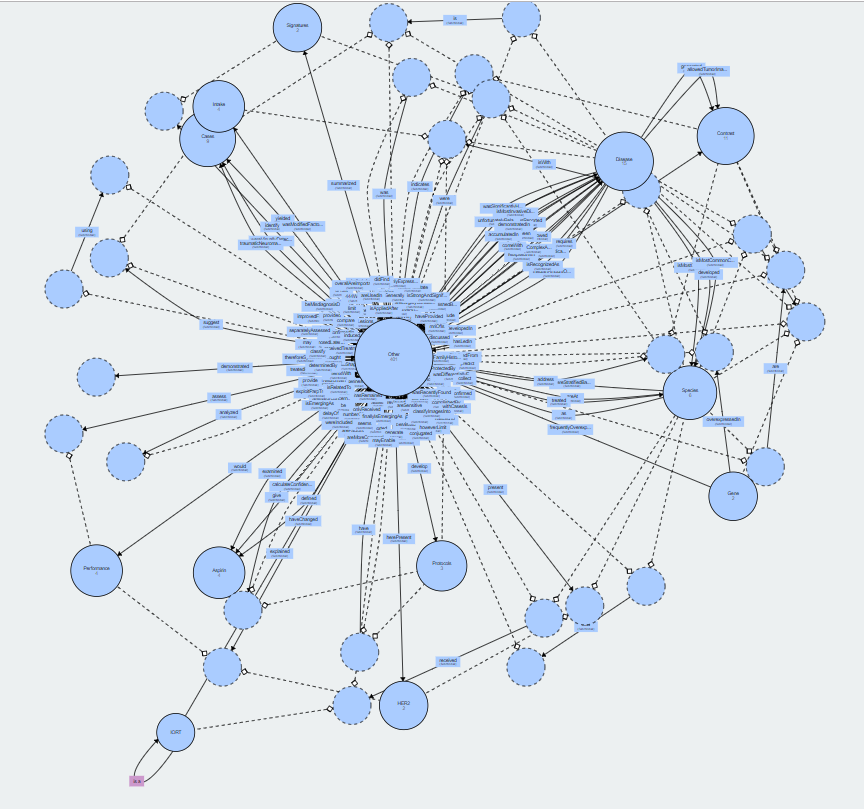


In the construction of this ontology subjects and objects are categorized solely into one of the 5 bioNLP medical word categories or as ‘other’, and predicates have domains and ranges specified as one either one of the 5 bioNLP medical word categories, or ‘other’.

The structure of this ontology is very basic and in version 2 it is updated so that individuals are one of the 5 bioNLP medical words or one of the top 33 tf idf terms or ‘other’.

Version 2 Statistics:

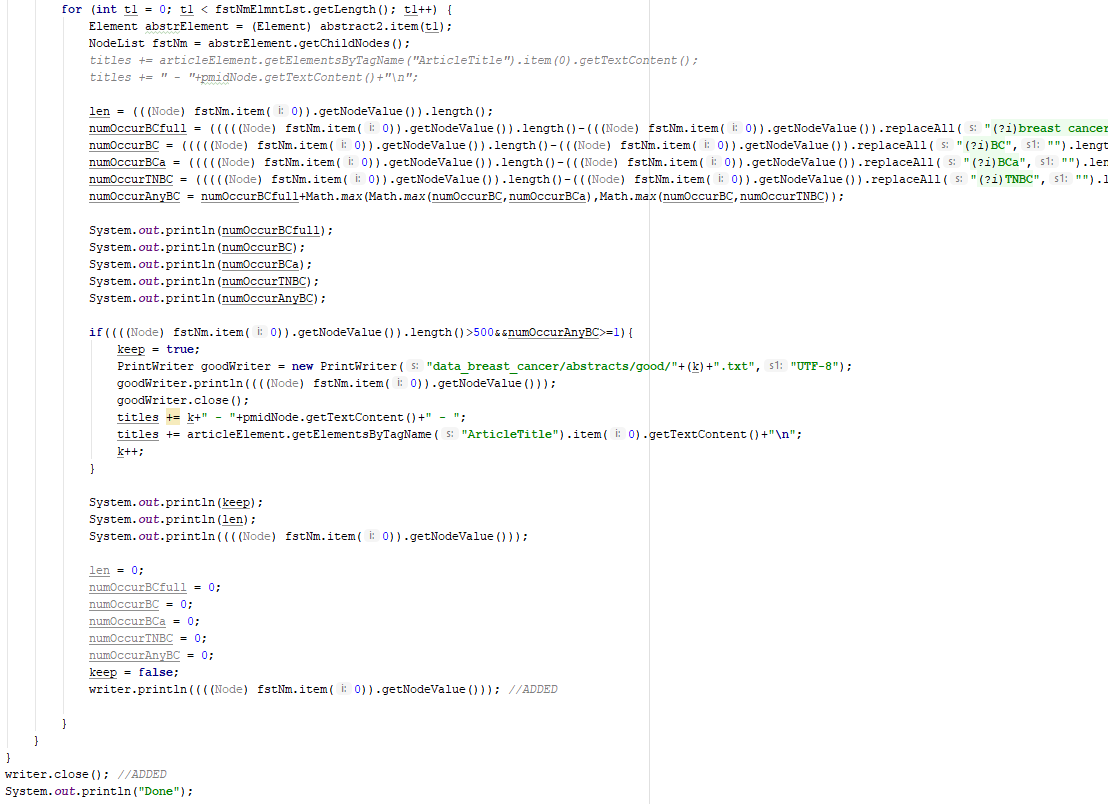
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Triplets | Subjects | Objects | Predicates | Classes | Subclasses | Individuals | Object Properties |
| 384 | 227 | 326 | 158 | 38 | 0 | 519 | 205 |



Version 2 is visualized above. This more sophisticated method for defining the structure of the ontology does yield additional classes (some of which--e.g. Contrast--are desirable, and some of which--e.g. Aspirin--are less desirable). However the structure of the resulting ontology is messy, and there are many, many individuals still within the ‘other’ class. Notably the method of defining structure is not substantially refined between this and version 3 (constructed for section 2), thus the structure of the Version 3 ontology is similarly messy, and also has many, many individuals still within the ‘other’ class.

Another notable feature of Version 2 is that there are fewer predicates than their are final object properties. This appears to be due to domain/range mapping of predicates. That is, some predicates have more than one pair of domain and range for which they are valid and thus some predicates yield multiple object properties.

**2) Create a single ontology from increased number of abstracts:**

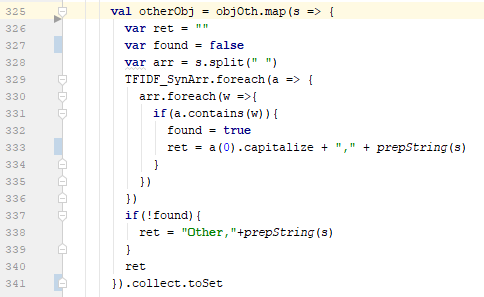
The first step in the process of constructing an ontology from additional abstracts was collecting additional abstracts. When attempting to get just 50 abstracts I noticed that the abstracts which were included in construction of the previous ontologies were not included. Furthermore all of the abstracts found were significantly more recent than the previous set of abstracts. Based on this information I decided to increase the pool of abstracts until all abstracts used in construction of the previous ontologies were included. This occured when the initial query of abstracts was increased to 750 (this speaks to the speed at which new research is being done on this topic and the potential benefit of programmatically parsing the new research). Additionally, I wrote new code to filter the abstracts. The criteria chosen were that abstracts used in the construction of this ontology should be at least 500 characters, and should include the word ‘Breast Cancer’ or some synonym, where synonyms include ‘Triple Negative Breast Cancer’, ‘TNBC’, ‘BC’, and ‘BCa’. Below is a screenshot of the code which does this filtering. 

Additionally these lines also construct a list of “<Doc number> - <PubMed ID> - <Abstract Title>” which will be used for step 3 of this lab. This list is output into a document so that it can be used to look up PubMed ID and Abstract title for a given entity or predicate.

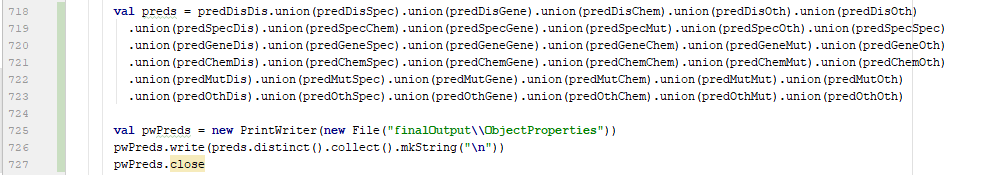
Once abstracts were collected, as in version 2, TF-IDF of subjects and objects was done in order to get candidates for grouping by top TF-IDF terms. This time the top 150 TF-IDF terms were identified and I manually went through and removed TF-IDF terms which already appeared in medical classes (i.e. ‘Cancer’ fell under ‘Disease’ and should not be included as a TF-IDF based class). With manually filtered list of top TF-IDF terms, Synonyms for each were found using the same script as in version 2. At this point I attempted to construct an ontology with this data, however, encountered several issues with special characters appearing in these documents. In order to solve these issues the following lines of code were used to clean-up the abstract text. This code comes very early on in the pre-processing of the data for ontology

.construction. As shown, many Greek characters as well as a few math symbols were replaced with english equivalents. As a special example the ‘greater than’ symbol is very important to replace because it interacts with the XML formatting of the final .OWL files output through ontology construction.

Since I did not provide an explanation for the method by which the pre-processing program does grouping by TF-IDF terms in the explanation of version 2, I will do so here. First the TF-IDF terms, and Synonyms thereof, are read into an RDD of Array[String] where each array is a list of synonymous terms, and then simply collected into an Array[Array[String]]. After grouping by medical terms, subjects and objects are grouped into these TF-IDF classes by, per subject/object, searching the array to see if any of the words in a given subject or object are contained within the array. If a match is found, that TF-IDF term or synonym is used as the class for that subject/object. If no match is found, ‘Other’ is used as the class for that TF-IDF term. The code on the next page shows how this process is performed on objects.



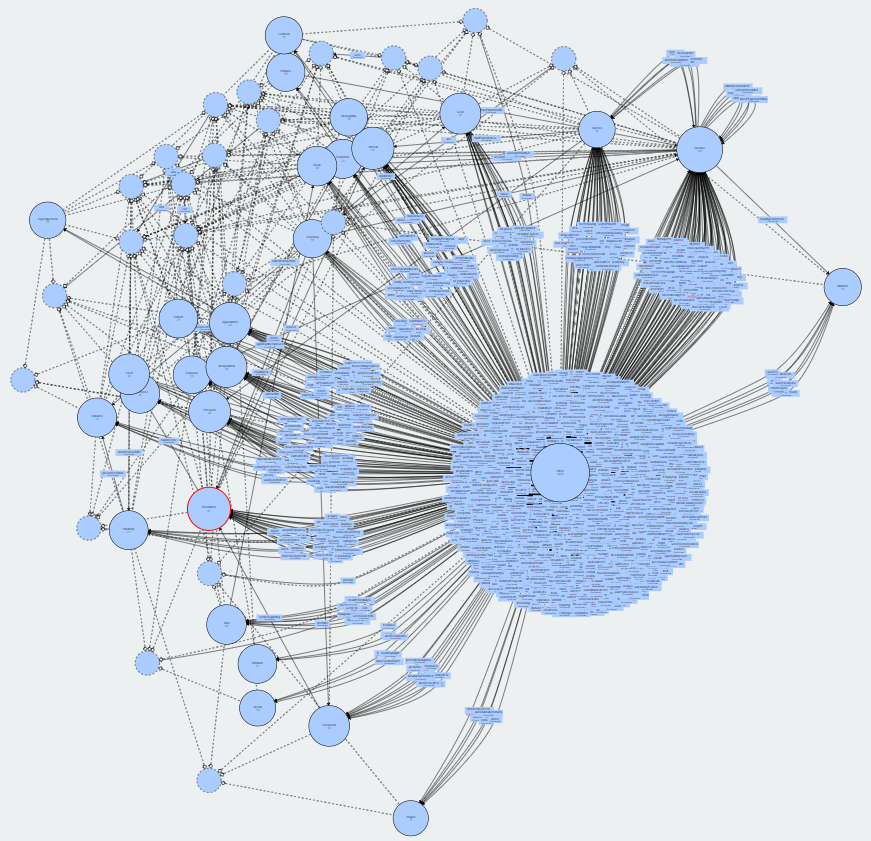
One final change made to the data-pre-processing script was to output the data in the same format as it is taken in by the ontology construction script. For version 2 outputs are produced separately (e.g. objects under the ‘disease’ class had the line which will be used to add them as individuals output into a separate file than objects under the ‘gene’ class and, for that matter, subjects under the ‘disease’ class were also output into a separate file). Outputting all individuals together, and all object properties (i.e. predicates) together was simply done by union of each of the relevant previous output sets (which previously were manually combined). Below is an example of this unioning process for ObjectProperties.



These changes made to the method performed to construct version 2, were sufficient to construct version 3 of the Breast Cancer Diagnosis ontology from >50 abstracts. Below are the data statistics about the constructed ontology.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Triplets | Subjects | Predicates | Objects | Classes | Sub-  Classes | Individuals | Object Properties |
| 4066 | 1947 | 1279 | 2985 | 366 | 0 | 4552 | 1275 |

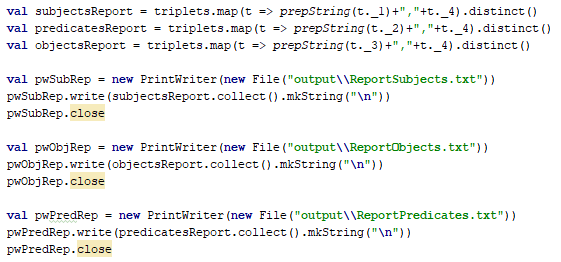
Unfortunately, there are still very many individuals (3481, to be specify) which are being categorized into the ‘Other’ class, meaning only about a thousand are being caught by grouping by Top TF-IDF terms. Additionally, as mentioned at the end of part 1, the structure of this ontology still doesn’t include subclasses and is very messy. Below is a visualization of the constructed ontology:



As you can see it is much denser than version 2, and although the ‘Other’ class is overused, there are still some positive features of this ontology which will be discussed in part 4 of this document.

**3) Report the Origins of entities and predicates:**

Updating the pre-processing script to enable us to trace back to the origins of entities and predicates was fairly straightforward. As show in the 2nd code screenshot from part 2 (on page 4 of this document) the document name of the abstract (e.g. ‘46.txt’) can be passed along with each triplet without issue. Then this information can be output along with the corresponding triplet piece to identify where that entity or predicate came from. Below is a code snippet which outputs origin information. Outputting before any further processing has the advantage that we



get to see each of the places where an entity or predicate came from (which would not be possible if we did this at the end of the process, after removing duplicates, by entity/predicate name). The below tables contain examples of entities and predicates with their origins:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Class | Entity (Object) | Doc | PubMedID (s) | Article Title |
| Disease | fibroadenoma | 194.txt | 30202040 | Radiomics of US texture features in differential diagnosis between triple-negative breast cancer and fibroadenoma. |
| Disease & Carcinoma | carcinoma | 73.txt, 90.txt | 30328524;  30304305 | Cavernous malformations are rare sequelae of stereotactic radiosurgery for brain metastases.;  Erysipelatoid Carcinoma. |
| Neuropathy | lesion | 121.txt, 166.txt | 30275864;  30232767 | Extraocular Muscles Involvement as the Initial Presentation in Metastatic Breast Cancer.;  Radiological review of skull lesions. |
| Appearance | accuracyOfLesionDetection | 60.txt | 30345046 | Usefulness of multi-parametric MRI for diagnosis of invasive urothelial cancer: Case reports of bladder and ureteral cancer. |
| Other | need | 38.txt,  241.txt | 30363150;  30167086 | Two cases of mimics of bone metastasis in breast cancer.;  Inborn-like errors of metabolism are determinants of breast cancer risk, clinical response and survival: a study of human biochemical individuality. |

This first table shows Objects and their origins, the following table shows Subjects with origins:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Class | Entity (Subject) | Doc. | PubMedID | Abstract Title (or example) |
| Gene | TP53 | 110.txt | 30287823 | Germline pathogenic variants of 11 breast cancer genes in 7,051 Japanese patients and 11,241 controls. |
| Disease & Carcinoma | carcinoma | 7.txt, 67.txt, 236.txt | 30400595;  30338173;  30172235 | Breast Cancer Estimate Modeling via PDE Thermal Analysis Algorithms.; ...  (Only 1st Abs. Title included) |
| Neuropathy | lesion | 6.txt, 73.txt, 90.txt, 220.txt | 30402234;  30328524;  30304305;  30183988 | Gastric outlet obstruction caused by metastatic tumor of the stomach originating from primary breast cancer: A case report.; ...  (Only 1st Abs. Title included) |
| Obstacle | blockade | 211.txt | 30191237 | An all-in-one nanoparticle (AION) contrast agent for breast cancer screening with DEM-CT-MRI-NIRF imaging. |
| Other | need | 23.txt, 120.txt | 30377575;  30276443 | Cancer information needs according to cancer type: A content analysis of data from Japan's largest cancer information website.;...  (Only 1st Abs. Title included) |

The third table (below) reports examples of predicates and their origins:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Domain | Range | Predicate | Doc. | PubMedID | Abstract Title (or example) |
| Disease | Disease | isSometimesMistakenFor | 194.txt | 30202040 | Radiomics of US texture features in differential diagnosis between triple-negative breast cancer and fibroadenoma. |
| Chemical | Gene | inhibitorOf | 105.txt | 30290049 | Forkhead box C1 boosts triple-negative breast cancer metastasis through activating the transcription of chemokine receptor-4. |
| Disease | Species | isMostCommonCancerAmong | 136.txt, 225.txt | 30259416;  .  .  .  30179186 | Emerging ways to treat breast cancer: will promises be met?;  Advantages of breast cancer visualization and characterization using synchrotron radiation phase-contrast tomography. |
| Markers | Other | wereElevatedIn | 48.txt | 30357709 | B Cell-Attracting Chemokine-1 and Progranulin in Bronchoalveolar Lavage Fluid of Patients with Advanced Non-small Cell Lung Cancer: New Prognostic Factors. |
| <many> | <many> | is | 1.txt,  4.txt,  9.txt,  10.txt,  11.txt,  12.txt,  14.txt,  19.txt,  20.txt, ... | 30407215;30403909;  3039428;  ... | Factors associated with locoregional and metastatic breast cancer at diagnosis in a Southern Portuguese registry in the period 2005-2012.; ...  (Only 1st Abs.Title) |

Some notable characteristics of the reported examples will be expanded on in part 4 below.

**4) Report meaningful Common or Unique entities and predicate:**

As shown in the last entry of the example predicates with their origin’s table (above) ‘is’ is the most commonly occuring predicate, coming from 155 of the 246 abstracts used to construct this ontology. Additionally ‘is’ is an example of a predicate which has multiple domains and ranges. Specifically there are unique 34 combinations of <domain, range> for ‘is’.

The previous tables also contain many examples of unique entities and predicates. ‘fibroadenoma’, ‘accuracyOfLesionDetection’, ‘TP53’, and ‘blockade’ are unique entities in that they each come from only one document in the corpus. Similarly ‘isSometimesMistakenFor’, ‘inhibitorOf’, and ‘wereElevatedIn’ are unique predicates. Each of these unique entities and predicates appears to add meaningful information to the ontology based on my knowledge of breast cancer from working on this project.

Another category of interesting entities is entities which fall under multiple classes. There are <20 examples of entities in this category and most are categorized into the ‘Other’ class as well as a more meaningful class. ‘carcinoma’ (from the above tables) is an example of an entity in this category, belonging both to Disease and Carcinoma (it is the only example for which ‘Other’ is not one of the classes into which the entity is categorized). While it is desirable for the ontology construction method to yield entities in this category, the example of ‘carcinoma’ is actually undesirable as ‘carcinoma’ should probably be categorized as a ‘Disease’ (as it is) and should not itself be a Class. This simply results from the TF-IDF+filtering which was done to obtain additional classes for grouping, and will be rectified in later versions of this breast cancer diagnosis ontology.

One final aspect of the data I would like to explicate about is the class structure which has some good features and some bad features. The most glaring bad feature is that there are currently still no subclasses. A method for rectifying this would be to group by TF-IDF term in two rounds. In the first round subclasses will be created based on exact matching of entities with top TF-IDF terms. In the second round, those subclasses would be populated in individuals based on looser matching criteria (e.g. synonym based matching, and partial TF-IDF term matching). Having a subclass structure would provide the benefit of more logical ordering of entities and would solve the issue with ‘carcinoma’ presented above (i.e. ‘carcinoma’ would be a subclass of disease, and other entities currently belonging to the ‘carcinoma’ class would be individuals under that subclass).

The most notable good feature of the class structure is the inclusion of many meaningful classes and relevant individuals belonging to those classes. For example, while it is not necessarily ideal to include in a Diagnosis focused ontology, it is a positive sign that ‘Drugs’ is a class based on top TF-IDF terms, and many of the individuals which fall under the ‘Drugs’ class are promising. Below is a list of individuals belonging to the ‘Drugs’ class:

Drugs,cardiotoxicityOfHer2-targetedDrugs

Drugs,chemotherapyDrugs

Drugs,conventionalChemotherapyDrugs

Drugs,developmentOfEffectiveDrugsForTreatmentOfHormone-dependentBreastCancer

Drugs,developmentOfNewDrugs

Drugs,drugs

Drugs,resistanceToAntimitoticDrugs

Drugs,usedDrugs

Two to four of these individuals seem as though they do not necessarily belong in the drug class, namely cardiotoxicityOfHer2-targetedDrugs and resistanceToAntimitoticDrugs seem as though the primary entity is not a drug, but ‘cardiotoxicity’ in the one case and ‘resistance’ in the other. The two individuals which begin with ‘development’ seem as though ‘development’ is the primary entity, rather than drug, but perhaps should still fall under the ‘Drugs’ class.

Other desirable classes include ‘contraception’, ‘complications’, ‘document’, ‘probe’, ‘provisions’, ‘quantification’ and many others. Below are some examples of individuals which properly belong to these desirable classes.

Contraception,contraceptionUse

Contraception,hormonalContraceptionUse

Complications,lateComplications

Complications,occasionalComplications

Complications,acuteComplications

Complications,complications

Document,nomogram

Document,model

Document,form

Document,papers

Document,content

Document,copyright

Document,document

Document,literature

Probe,nirProbe

Probe,stressProbe

Provision,exercise

Provision,follow-up

Quantification,multidimensionalScaling

Quantification,observation

Quantification,quantification

These classes are, of course, not perfect and below are a couple example of individuals which should probably not belong to these classes:

Document,questionnaireBasedOnAnderson'SdelayModel

Quantification,whole-tumorSamplingOfDiseaseBurden

Thus far my best plan for removing these less desired indivduals from these classes is to ensure that they end up in other classes by increasing number of the TF-IDF terms considered and to limit inclusion in a given class to only include entities with a few words which are not relevant to that class. Testing will be required to determine what ‘a few’ means in this case in order to achieve the best results.

Finally there are some Classes which either themselves seem as though they should not be classes, or seem as though they should not be classes based on the individuals which belong to them. Examples of such classes include ‘Appearance’, ‘Collision’, and ‘Behaviors’. The ‘appearance’ class is an issue for both mentioned reasons, the former because it is very general and seems likely to include strange individuals based on synonyms, and because it specifically does include strange in irrelevant individuals, for example:

Appearance,accuracy

Appearance,surgicalAccuracy

Appearance,theirAbility

The ‘Behaviors’ class is almost a third category of reason for which it should perhaps not be a class, specifically because it only includes four individuals:

Behaviours,changesInHealthBehavioursAfterBreastCancerDiagnosis

Behaviours,healthBehaviours

Behaviours,positiveChangesInTheirHealthBehaviours

Behaviours,theirHealthBehaviours

Some of these individuals are fine, ‘healthBehaviors’ for example seems appropriate, however does not seem to warrant a class of its own, and the others are redundant, and somewhat complicated.

The ‘Collision’ class is the worst of these three examples because it contains individuals which it is understandable that were added based on synonyms to the TF-IDF term ‘collision’ but which are problematic in that they use a different sense of the synonym which was found:

Collision,typeOfCollisionTumor

Collision,potentialImpact

Collision,familyImpact

Collision,recurrentWords

The only listed example which seems to refer to the appropriate sense of meaning of the word collision is ‘typeOfCollisionTumor’. However ‘impact’ in both individuals in which it is included carries a different sense of meaning than the word ‘impact’ for which ‘collision’ is a synonym.

In closing there are several good features of the version 3 ontology and also several features which ought to be changed. Future versions of this ontology will attempt to improve on the class structure, categorize more individuals into meaningful classes (so that fewer fall into the ‘Other’ category) and generally clean-up other mentioned issues with version 3.