Greg Brown Class ID: 1

# 1. Ontology Description

## 1. Determine the domain and scope of the ontology.

The ontology (and various versions thereof) which I have constructed aim to cover the domain of Breast Cancer Diagnosis. More broadly the ontologies that our team has been working on aim to cover the more general domain of the disease Breast Cancer. Ideally the ontologies we have constructed could be used in Clinical Decision Support software to aid decision making surrounding diagnosis and treatment of breast cancer, and could help doctors answer questions like what stage of breast cancer is characterized by particular disease presentation (symptoms) or what treatment would be appropriate for a particular case of breast cancer. However, in their current state these ontologies are unlikely to provide useful answers to such questions.

In terms of scope the ontologies our team has constructed, and more specifically the ontologies that I have constructed, are intended to be limited to Breast Cancer, and it is not expected that they provide useful answers to questions about other Cancers or other Diseases. Furthermore the scope is limited to medical questions.

## 2. Consider reusing existing ontologies.

I have explored the NCIT (National Cancer Institute Thesaurus) ontology with an eye to structural similarities between it and the ontologies which I have constructed. Additionally our team has reviewed the strategies for construction of the BCS7 and BCS8 ontologies, corresponding to the AJCC 7th edition and 8th addition respectively. Early on we considered using these texts (AJCCs) as additional sources for our ontology, but determined it infeasible for our current project. We did not explicitly consider using entities or structural relationship from these ontologies to provide structure and entities to the ontologies we constructed, but this had more to do with our lack of understanding of the ontology construction process at the outset of this project and our understanding of the aim of this project, which we took to be the automatic generation of Class-Subclass-Individual structure of our ontologies from unstructured data, which we did not consider these ontologies to be. Retrospectively, using structured data to lend structure to our data is an interesting prospect.

## 3. Enumerate important terms.

For the scope that I specified (in 1.) the following terms have shown up as classes which I have verified either by comparison with the NCIT ontology or in conversation with Dr. Ann Lottes (my mother; they are listed in a table to conserve space):

|  |  |  |  |
| --- | --- | --- | --- |
| Activites  Autoantibodies  Behaviors  Biopsy  Cell  Chemical  Contrast  Disease | Exposure  Factor  Gene  Imaging  Marker  Mechanism  Mutation  Neuromas | Phenotype  Presentations  Protocols  Quantification  Radiotherapy  Resistance  Risk  Signature | Species  Survival  Symptom  Targets  Technique  Toxicity  Treatment  Tumor |

Among these there are the following overlaps:

* ‘Behaviors’ and ‘Technique’ could be categorized under ‘Activities’
* ‘Contrast’ could be categorized under ‘Imaging’
* ‘Factor’ might be best categorized under ‘Marker’ alongside ‘Signature’, or alongside ‘Symptom’ and under ‘Presentations’, or alongside ‘Phenotype’ and ‘Exposure’ under ‘Risk’
* ‘Biopsy’ could be categorized under ‘Activities’ alongside ‘Technique’
* ‘Radiotherapy’ could be categorized under ‘Treatment’
* ‘Toxicity’ and ‘Survival’ could be categorized under ‘Risk’
* ‘Mutation’ could be categorized under ‘Disease’
* ‘Tumor’ could be categorized under ‘Mutation’ or ‘Disease’

These changes would leave me with the following important terms:

|  |  |  |  |
| --- | --- | --- | --- |
| Activites  Autoantibodies  Biopsy  Cell  Chemical | Disease  Gene  Imaging  Marker  Mechanism | Mutation  Presentations  Protocols  Quantification  Resistance | Risk  Signature  Species  Targets  Treatment |

## 4. Define the classes & class hierarchy.

Largely definition of classes is established previously (in 3.), however the following are a few classes which do not appear in my ontology, but appear in the NCIT ontology, and which my ontology might benefit from including:

* Anatomic Structure, System, or Substance
* Biochemical Pathway
* Biological Process
* Drug
* Food
* Biomedical Material
* Manufactured Object
* Molecular Abnormality

Additionally the ontologies I have constructed have no subclass structure (thought the explanation given previously (in 3.) give some indication of my inclinations about certain subclasses. There are further subclasses which could arise from sets of individuals (e.g. ‘cancer’ as a ‘subclass’ of ‘disease’ could include ‘breastCancer’, ‘lungCancer’, etc which currently show up alongside ‘cancer’ under ‘disease’. This kind of subclass structure, could be incorporated into the ontologies I have made in a few ways I can think of, but I am not sure which would be best: simple String comparison of terms within subclass choosing the most general tem (probably the shortest) as the subclass and equivalent terms as individuals under the subclass (this would be simplest); performing two rounds of TF-IDF in order to find classes and the subclasses; performing two rounds of LDA in order to find classes and subclasses; a combination of TF-IDF and LDA.

## 5. Define the properties of classes.

In the ontologies I have constructed the properties of the classes (or rather their individuals) are generated automatically from the predicates of <subject, predicate, object> triplets, using the classes of the subjects and objects as domain and range. These properties are also referred to as ‘slots’ (as I understand it) in the referenced literature from Stanford, with ‘intrinsic’ properties equivalent to our object properties and ‘extrinsic’ properties as equivalent to our ‘data’ properties. I have not defined any data properties, nor are data properties defined in the other ontologies constructed by our team. Below I will discuss how the “facets” of our object properties are set, by which I understand the Stanford literature to mean the entities connected by the Object Properties (as well as what the ‘values’ of the data properties would be if we had data properties).

## 6. Define the facets of the slots.

Defining the connections between the entities (individuals/classes) via the object properties in the ontologies I’ve constructed is done in the final pre-processing step, after individuals are slotted under classes. The domain is specified as the class to which the subject in a <subject, predicate, object> triplet belongs for a predicate (i.e. object property, in question), and similarly the range is specified as the class to which the object in the same triplet belongs. As are result a given object property may have multiple domains if the relevant predicate appears in more than one triplet and the relevant subjects/objects belong to different classes respectively.

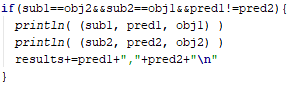
## 7. Create instances.

The instances are the individuals which are the subjects and objects from the triplets from the sentences of the documents in the corpus.

# 2. Characterize the triplets based on the following Rules

## a. Inverse Of

The Inverse-Of relation holds of a pair of predicates (or properties if you prefer), P1 and P2, when there exists some pair of triplets, T1 and T2 in the set of triplets, such that T1 = <S, P1, O> and T2 = <O, P2, S> that is, the subject, S, of the first triplet is the object of the second triplet and the object, O, of the first triplet is the subject of the second triplet. For example if we have the triplets <Greg, hasParent, Scott> and <Scott, hasChild, Greg>, then ‘hasParent’ and ‘hasChild’ are considered inverses of one another. Below is the primary section of code used to check this property in the script I wrote to find such pairs of relations:



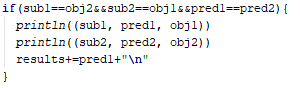
This code section comes after reading in the set of triplets to two arrays, and comparing each triplet in the first array to each triplet in the second array, looking for the subject of the first triplet to equal the object of the second triplet, the object of the first triplet to equal the subject of the second triplet and the predicates to not be equal. If a pair of triplets matching this criteria is found, those predicates are added to a String which accumulates relations (predicates) which function as inverses of one another. Later this ‘results’ String is split into an array, duplicates are removed, then the array is made back into a string and output to a text file. The below table contains the two sets of triplets which were found to be inverses of one another:

|  |  |
| --- | --- |
| **Relations which are Inverses Of one another** | **Triplets yielding Inverse Relations** |
| isMostCommonCancerAmong,isWith  isWith,isMostCommonCancerAmong | (breastCancer,isMostCommonCancerAmong,women)  (women,isWith,breastCancer) |
| isWith,isMostCommonCancerIn  isMostCommonCancerIn,isWith | (breastCancer,isMostCommonCancerIn,women)  (women,isWith,breastCancer) |

As will be discussed further in section 3 this type of pair of relation does not occur frequently in the corpus of documents. It is noteworthy that this script was run on a the triplets resulting from a set of 187 abstracts which were used to construct Version 4 of the Breast Cancer Diagnosis ontology.

## b. Symmetric Property

A relation is said to be Symmetric if a predicate P is found in two triplets, T1 and T2, such that T1 = <S,P,O> and T2 = <O, P, S>, that is the subject of the first triplet is the object of the second triplet, the object of the first triplet is the subject of the second triplet, and the predicate for each triplet is the same. The below code snippet is used to check this property in the same way as for the inverse of property, i.e. after loading the set of triplets into two arrays and comparing each triplet in the first array to each triplet in the second array:



Notably this condition is very similar to the conditions Inverse Of Relations, the only difference being that the predicate should be the same in both triplets for the Symmetric property to hold of that predicate. The outputting of the relations fulfilling these conditions is done in the same way as for the Inverse Of Relations as well as each other type of relation listed subsequently. Below is the only example of a such a relation found within this corpus:

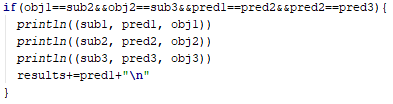
|  |  |
| --- | --- |
| **Symmetric Relations** | **Triplets yielding Symmetric Relations** |
| explain | (heterogeneity,explain,heterogeneity) |

As with the Inverse Of pairs, relations meeting the conditions to be considered Symmetric are very rare. Also, the only triplet which was found to meet this Symmetric relation is also Reflexive. This is to be expected as an Reflexive Relation should also be a Symmetric Relation, and for that matter a Transitive Relation.

## c. Transitive Property

The main difference in the script used to find Transitive Relations in addition to the conditions which must be met to consider a relation transitive is that we must load the triplets into three arrays and compare each set of three triplets arising from those three sets of triplets. This exponentially increases the computation time.

As for the conditions which which a relation must meet in order to be considered Transitive the relation must be a predicate P, such that there are three triplets, T1, T2, and T3, such that T1 = <S1, P, O1>, T2 = <S2, P, O2> and T3 = <S3, P, O3> where O1 = S2 and O2 = S3 and the the predicate for each triplet is the same. The following code section is the part of the script I used to find relations meeting these conditions:



And the following table contains all the transitive relations found by my script:

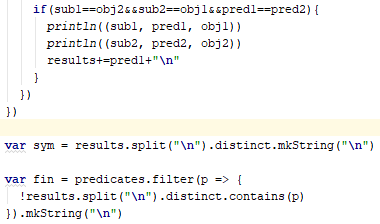
|  |  |
| --- | --- |
| **Transitive Relations** | **Triplets yielding Transitive Relations** |
| explain | (heterogeneity,explain,heterogeneity)  (heterogeneity,explain,heterogeneity)  (heterogeneity,explain,heterogeneity) |
| isIn | (Delays,isIn,detection)  (detection,isIn,women)  (women,isIn,generalPopulation),...  <multiple sets of triplets yielded ‘isIn’ relation> |
| is | (carcinoma,is,breastCancer)  (breastCancer,is,cancer)  (cancer,is,heterogeneous),...  <multiple sets of triplets yielded ‘is’ relation> |

## d. Property Chain Axiom

I did not successfully implement the Property Chain checking

## e. Asymmetric Property

The Asymmetric Relations should be precisely those which are not Symmetric:



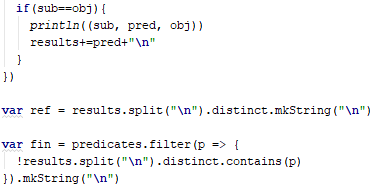
As shown the symmetric property is found by determining which predicates are symmetric and taking the total set of predicates less the symmetric predicates.

|  |
| --- |
| **Asymmetric Relations (examples)** |
| alter |
| furthermoreWasAssociatedWith |
| are |

Since only one relation was found to be symmetric the remaining 952 predicates were asymmetric, hence only a few examples are shown in the table above. Also, because of they are defined as the complement of the symmetric relations, the script used did not output the triplet which yielded the asymmetric relations.

## f. Irreflexive Property

As I have not yet shown a method of finding reflexive relations, it is worth briefly describing the conditions which must be met of a given relation in order to consider it reflexive. We will then use the fact that the Irreflexive Relations are simply those relations which are not reflexive to find the Irreflexive Relations. For a relation to be considered reflexive we need only find a triplet T such that T = <S, P, S>, i.e. the subject and object of that triplet are the same. For a triplet meeting this criteria the predicate, P, of that triplet is a reflexive relation. Below is the code section used to find the Reflexive relations and subsequently the Irreflexive Relations:



Similar to the Asymmetric Relations, the Irreflexive relations are defined as the complement of the reflexive relations, hence, as with the Asymmetric Relations we will maintain a list of just the predicates, and once we find the reflexive relations we will simply remove those relations for the list of all relations to find the Irreflexive ones.

|  |
| --- |
| **Irreflexive Relations** |
| received |
| isIn |
| verified |

Incidentally, the only Symmetric Relation which was found is identical to the only Reflexive Relation which was found hence the list of Irreflexive Relations is identical to the list of Asymmetric Relations, containing the same 952 relations. Thus rather than reproducing the same few relations in the above table the next few are included.

# 3. Record unique features noted from your ontology

As briefly mentioned earlier the InverseOf, Symmetric, and Transitive relations are relatively rare, making up less than 1% of the relations in the corpus each, thus while the Symmetric relation found is the only one which is truly unique in the sense that it is the only relation which has that property, the InverseOf relations and Transitive Relations are substantially rare and worth discussion. Unfortunately very few of these relations are meaningful in our corpus and thus probably should not be implemented within the final ontology. The two InverseOf Relations seem to stem from the same sentences and should likely only be considered one relation by excluding one from the corpus. Furthermore while the ‘hasChild’ and ‘hasParent’ relationed mentioned as an example of the InverseOf property of Relations does follow a natural language understanding of what an inverse relation should be ‘isWith’ and ‘isMostCommonCancerIn’ do not naturally sound like or seem to be inverses of one another, ‘isWith’ being a very broad and general relation, whereas ‘isMostCommonCancerIn’ being very specific.

Similarly the triplet which yields the on Symmetric relation does not convey hardly any information, and is actually confusing. To say that ‘heterogeneity explain(s) heterogeneity’ seems to leave out some information; for example it might make sense if we understood it as ‘heterogeneity of tumors explain(s) heterogeneity of outcomes’ or ‘... heterogeneity of appropriate treatments’ might carries much more meaning but also loses the property of being Symmetric. In fact, based on work done previously we can find that the sentence which produced this triplet is

“The intrinsic molecular subtypes can explain the intertumoral heterogeneity and the cancer stem cell (CSC) hypothesis can explain the intratumoral heterogeneity of this kind of tumor”

from the abstract of the article “A Multilevel Examination of Health Disparity: The Roles of Policy, Neighborhood Context, Patient Resources and Healthcare Facilities in Stage at Diagnosis”. This sentence does seems likely to carry more information than the triplet representing it--although that information is totally lost on me--and almost seems to maintain the Symmetric property of the relation which is the predicate in this sentence, but in my vague understanding does seem to fail to be Symmetric.

The best quasi-unique relations, by way of these properties, are the Asymmetric Transitive properties ‘is’ and ‘isIn’. Furthermore these relations were found to be Transitive by multiple sets of triplets, lending credibility to the notion that they truly ought to be Transitive. Furthermore, in natural language these relations are often transitive, for example ‘green IS a color’ and ‘a color IS a range of light wavelengths’, and furthermore ‘green IS a range of light wavelengths’. Likewise, to lend some sense of the natural language transitivity of the IS\_IN relation consider that if ‘a tomato slice IS\_IN a sandwich’ and ‘a sandwich IS\_IN a bag’ (and we are referring to the same sandwich) then ‘a tomato slice IS\_IN a bag’.

I will add that relations fulfilling the Property Chain Axiom, if there were any, may well also have been better examples of unique and meaningful relations as the conditions of the Property Chain Axiom is similar to those of Transitive relations (as well as Inverse Of relations, which were at least more meaningful on-sight than the one Symmetric relation found). However as I was unable to implement a script to find relations which meet those conditions thus far, I do not yet know.

One final not is that Version 4 of the ontology does seem slightly better than previous versions, though only slightly, and the improvement mostly owing to an improved mechanism of filtering abstracts for use in ontology construction which I implemented as well as slightly improved search terms syntax found through exploration of PubMed mesh-terms. Still most of the entities and relations in Version 4 of this ontology confusing or meaningless, but there are no entities or relations which are as misleading as the worst examples from Version 3.