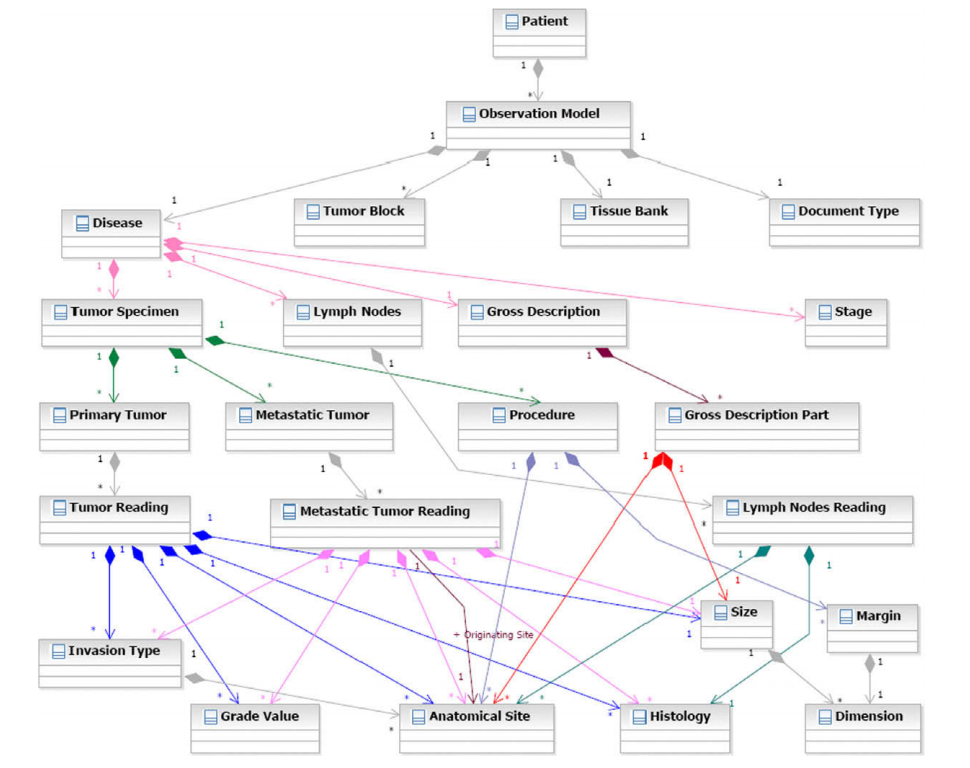
**Automatically extracting cancer disease characteristics from pathology reports into a Disease Knowledge Representation Model**

* Ingestion and tokenization: inserting annotations (extracting implicit meaning from a document)
* Sentence annotations: determining sentence boundaries
* POS tagger and shallow parser: statistical POS tagger (with a manually annotated corpus augmented) plus POS tags from a dictionary of terms applicable within pathology reports. The shallow parser identifies (1) noun phrases, (2) noun phrase lists and (3) prepositional noun phrases ?? -> hierarchy (do POS for?)
* Concept identification: ConceptFinding maps textual mentions (from the text) to terminology concepts.

1. conceptMapper: creates candidate matches between concept structures based on a terminology and unstructured text

Leaf classes

1. conceptFilter: a set of rule-based annotators that filter out matches

* Pattern recognition and negation: regex in conjunction with a terminology discover textual mentions describing dimensions and sizes, dates, number of excised and positive lymph nodes as well as stage. Negation detector is used to detect negated phrases within a user-specified window.
* Relation identification: discover the relationship between leaf classes (e.g. histology, anatomical sites, size, grade) to populate container classes e.g. primary and metastatic tumor classes, the lymph node class and the gross description part class.
* First step: determine which section of a document should be considered for instantiating a container class.
* Second step: categorize certain classes based on multiple criteria. The categorization algorithm is based on a set of trigger phrases and the noun phrase hierarchy. The algorithm checks whether an appropriate trigger word is co-occurrent with the mention attribute of the class. ICD-O codes are used to categorize histologies and anatomical sites.
* Third step: co-referring (determine which mentions refer to each other)
* Fourth step: determine which instances of classes should be candidates for populating each of the container classes
* Fifth step: container classes are merged or split according to class-specific rules

Container classes

More specific details to instantiate the primary and metastatic tumor classes

* Determine tumor context TC (a user-defined portion of the document)
* Identify all non-negated histlogies winthin TC
* For all identified histologies, examine noun phrase (or noun phrase list) containing the histology for all occurrences of any of anatomical sites, grade values and sizes and associate them with the histology.
* Repeat the previous step for histologies missing one or more of these associations
* Tumor classes which have co-referenced histologies are merged into a single instance
* Instantiate the tumor classes based on the categorization of the histology and anatomical sites.

More specific details to instantiate the lymph nodes classes

* Only anatomical sites which have been categorized as lymph nodes (AS-L) are considered
* For each AS-L, the algorithm determines the histologies and lymph node expressions(LNE) co-occurring with the AS-L in the same sentence (or other sentences in the same section)
* A set of rule-based filters is applied to derive the associations, taking the categorization of histologies and anatomical sites into positive and negative classes into account.

More specific details to instantiate the gross description part class

* Consider the ParenthesesSeparatedNoun(PSN): a sequence of a noun phrase, a parenthesis, a noun phrase followed by another parenthesis. And the ParenthesesPhrase (PPH): an expression enclosed with matching parenthesis
* One or more gross description part classes are instantiated if anatomical site(s) and size expression(s) co-occur in the same syntactic structure?
* Basic approach to map entities to a standard terminology (Used in this paper) is to augment a standard terminology, e.g. with plural forms and synonyms (by using variations in punctuation, removing parenthesized expressions from a mention, removing stop-words or punctuations, using some common abbreviations or using adj forms of a word), and create the links between the text and the augmented terminology