SciBite – turning text into ontologies

Jane Lomax

The SciBite tool TERMite takes text, in any format, and extracts the scientific terminology using named entity recognition (NER). Powering TERMite are more than 80 biomedical vocabularies. These vocabularies are derived from publicly available ontologies and vocabularies where possible and include HGNC for human genes, MeSH, NCBI Taxonomy ontology, Human Phenotype Ontology (HPO) and the Gene Ontology (GO). SciBite vocabularies are optimised for text-mining by curation to add additional synonyms and using a rule-based expansion to generate lexical variants. Context-specific rules are also added to the vocabularies to take into account ambiguous phrases such as 'hedgehog' which can be either a gene or an organism. In addition, SciBite provides a suite of tools for editing, validation, compiling, regression testing these synonym-centric vocabularies.