# Annotation

## Services

* NCBO (bioportal’s annotation REST service)
  + Pros:
    - Little coding necessary, just utilize the REST API
    - Build on existing biotea code for annotation
    - Lots of ontologies available
  + Cons:
    - Available ontologies pretty much restricted to biomedical domain (genetics..)
    - Contains the APA ontology, but only in a flat format (needs 1:1 mapping)
    - No NLP (TODO verify this) done, relies on 1:1 mapping (no stemming, lemmatization etc.) -> lots of annotations are missed out because of this
* UMLS
  + Pros:
    - Existing API
    - NLP tools exist (TODO utilize how?)
  + Cons:
    - ?? Easy to use?
    - Also pretty much restricted to biomedical domain
    - No APA thesaurus
* MAUI (to utilize APA thesaurus in DFKI SKOS format with hierarchy)
  + (Just for APA thesaurus), and for other tasks than annotation (e.g. finding more keywords)
  + ?? Needs much configuration work / adaptation?

Ontologies

* Recommender results see excel sheet in this directory. TODO repeat recommendation: At least two documents were not in English and therefore have not been considered.
* Using the NCBO recommender service, the abstracts of all PsychOpen articles have been analyzed.

Results:

* + Thesauri that contain very general concepts, such as the NCIT and MESH,
  + On the other end of the spectrum very specific ontologies, that touch the realm of psychology, such as the ontology of Alzheimer’s disease, the suicide ontology
* TODO check UMLS Thesaurus for more ontologies.

Annotation with AO ontology

* The matches belonging to the same context must be captured in a list. Otherwise, we cannot query how the mapping has been performed.

## Implemented Annotators

* UMLSAnnotator
* NCBOAnnotator

### UMLSAnnotator

* Uses MetaMap 2014 server
  + The data/lexicon is a 2013 NLM release, relaxed model
    - NLM release includes licensed terminologies, including the APA thesaurus. For 2014, no windows model exists yet, so I took the 2013 one.
* Does not use optional Word Sense Disambiguation (getting a java heap space error. Could be sorted out, but did not bother yet)
* TODO describe MetaMap (NLP preprocessing steps including tagger, describe options)
* Options: default options.
* Model: relaxed model (more recall, but maybe weaker matches included than in strict model)
* RDFIzation
  + Currently,

### NCBOAnnotator

* Uses APA cluster ontologies plus APA full/aggregated ontology.

## Queries

* Most frequent terms across the entire database

select (sum(?occurrences) as ?count) ?body

where {

?qual rdf:type aot:ExactQualifier.

?qual ao:body ?body.

?qual biotea:occurrences ?occurrences.

}

GROUP BY ?body ORDER BY Desc (?count)

select ?body ?match ?para

where {

?qual rdf:type aot:ExactQualifier.

?qual ao:body ?body.

?qual ao:context ?context.

?context ao:exact ?match.

?context dcterms:references ?para.

}

Order the annotated concepts by frequency (occ. Within one paper) -> aggregation mistake because of context. Do not use!

prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

prefix owl: <http://www.w3.org/2002/07/owl#>

prefix xsd: <http://www.w3.org/2001/XMLSchema#>

prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>

prefix ao: <http://purl.org/ao/core/>

prefix aot: <http://purl.org/ao/types/>

prefix biotea: <http://www.biotea.ws/ontology/ao\_biotea.owl#>

Select distinct ?inTextCount ?body

where {

?term a aot:ExactQualifier .

?term ao:body ?body.

?term ao:hasTopic ?topic.

?term biotea:occurrences ?inTextCount.

?term ao:context ?context.

?context ao:onResource <http://www.zpid.de/resource/doi/10.5964/ejcop.v3i1.23>

}

ORDER BY DESC (?inTextCount )

prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

prefix owl: <http://www.w3.org/2002/07/owl#>

prefix xsd: <http://www.w3.org/2001/XMLSchema#>

prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>

Select ?inTextCount ?topic ?body

where {

?term a aot:ExactQualifier .

?term ao:body ?body.

?term ao:hasTopic ?topic.

?term biotea:occurrences ?inTextCount

}

ORDER BY DESC (?inTextCount )

Select ?inTextCount ?topic ?body

where {

?term a aot:ExactQualifier .

?term ao:body ?body.

?term ao:hasTopic ?topic. ?term biotea:occurrences ?inTextCount

# ORDER BY DESC (?inTextCount )

}

prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

prefix owl: <http://www.w3.org/2002/07/owl#>

prefix xsd: <http://www.w3.org/2001/XMLSchema#>

prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>

prefix ao: <http://purl.org/ao/core/>

prefix aot: <http://purl.org/ao/types/>

prefix biotea: <http://www.biotea.ws/ontology/ao\_biotea.owl#>

Select (sum(?inTextCount) AS ?count) ?body ?topic

where {

?term a aot:ExactQualifier .

?term ao:body ?body.

#?body <http://www.w3.org/2000/01/rdf-schema#label> ?label .

?term ao:hasTopic ?topic.

?term biotea:occurrences ?inTextCount.

?term ao:context ?context.

#?context ao:onResource <http://www.zpid.de/resource/doi/10.5964/ejcop.v3i1.23>

}

GROUP BY ?label ?body ?topic ORDER BY DESC (?count )

query w Resource

prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

prefix owl: <http://www.w3.org/2002/07/owl#>

prefix xsd: <http://www.w3.org/2001/XMLSchema#>

prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>

prefix ao: <http://purl.org/ao/core/>

prefix aot: <http://purl.org/ao/types/>

prefix biotea: <http://www.biotea.ws/ontology/ao\_biotea.owl#>

Select ?inTextCount ?body ?topic ?resource

where {

?term a aot:ExactQualifier .

?term ao:body ?body.

#?body <http://www.w3.org/2000/01/rdf-schema#label> ?label .

?term ao:hasTopic ?topic.

?term biotea:occurrences ?inTextCount.

?term ao:context ?context.

?context ao:onResource ?resource

#?context ao:onResource <http://www.zpid.de/resource/doi/10.5964/ejcop.v3i1.23>

}

GROUP BY ?inTextCount ?body ?resource ?topic ORDER BY DESC (?inTextCount)

Filter all papers on inTextCount

prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

prefix owl: <http://www.w3.org/2002/07/owl#>

prefix xsd: <http://www.w3.org/2001/XMLSchema#>

prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>

prefix ao: <http://purl.org/ao/core/>

prefix aot: <http://purl.org/ao/types/>

prefix biotea: <http://www.biotea.ws/ontology/ao\_biotea.owl#>

Select (sum(?inTextCount) AS ?count) ?body ?topic

#Select ?inTextCount ?body ?topic

where {

?anno a aot:ExactQualifier .

?anno ao:body ?body.

?anno ao:hasTopic ?topic.

?anno biotea:occurrences ?inTextCount.

#?anno ao:context ?context.

#?context ao:onResource <http://www.zpid.de/resource/doi/10.5964/ejcop.v3i1.23>

}

GROUP BY ?body ?topic ORDER BY DESC (?count )

* Quality of automatic annotation: E.g. issues retrieving the correct hierarchy: example: NCIT concepts “family”: 1) as an abstract concept”A taxonomic category between Order and Genus. It consists of a group of organisms among which the differences are quite minor, e.g. Equiidae - horses and their relatives.”
  + 2) A domestic group, or a number of domestic groups linked through descent (demonstrated or stipulated) from a common ancestor, marriage, or adoption.
  + Are both used in PsychOpen annotations, although it is questionable that the context is taken into account (? How does the algorithm work?)
  + Is it possible to automatically enhance annotation quality through the feature in the new REST Api: “Filter by semantic type”?
  + TODO find out difference between NCIT and NCI Metathesaurus
* NCIT browser: http://ncit.nci.nih.gov/ncitbrowser