Europe PMC Annotated Full-text Corpus

Author List and order To be Decided

## Abstract

Named entity recognition (NER) is a widely used text-mining and natural language (NLP) sub-task. Deep learning methods have superseded traditional dictionary and rule-based NER approaches in past years. A high-quality dataset is essential to take full advantage of the recent deep learning advancements. A number of gold standard corpus annotating biomedical entities in abstract exist but lacks full-text corpus. At Europe PMC, we aim to explore the benefits of deep learning methods for our gene/protein, disease and organism entity recognition and developed a manually annotated full-text corpus for these entities. Our corpus comprises 300 full-text open access research articles carefully selected from Europe PMC. Over 72K mentions of biomedical concepts are identified across about 114K sentences in the corpus.

## Background

The rapid growth in biological research publications makes tracking research trends and assimilating knowledge into reference databases a challenging and a time consuming task. However, recent advances in natural language processing are now making it possible to integrate sophisticated machine learning-based tools based on the research literature into the workflows of biological resources such as STRING[1], neXtProt[2] have embedded text mining processes in their data workflows to provide additional support and veracity to the hosted curated information. Furthermore, Text mining tools like PubTator[3] are being used in biomedical literature curation. Traditionally, BioNER and relationship extraction systems were using dictionary and rule-based approaches[4-7]. However, such systems are often context independent which leads to problems such as false positive recognition due to ambiguous phrases and missing recognition due to the use of non-standard vocabulary. New deep learning models have shown promising results to alleviate these challenges, but high quality, large and representative training data is critical for that. This article describes a set of manually annotated 300 hundred open access full-text articles with CC-By licence.

Several training sets are available to support the development of machine learning models. For instance, the BioCreative V CDR corpus (BC5CDR) consists of 1,500 PubMed abstracts annotated with chemicals, diseases and chemical-disease interactions[8]. The Collaborative Annotation of a Large Biomedical Corpus (CALBC) provides a large-scale biomedical corpus, automatically generated by harmonising annotations produced from different systems[9]. It contains concept annotations of genes, diseases and species across 150,000 abstracts. The GENIA corpus consists of 2,000 abstracts with human annotations of terminal concepts in the GENIA ontology[10]. Other datasets, such as JNLPBA, Yapex, NCBI disease, MedMentions and the CHEMDNER are also all based on PubMed abstracts[11-15].

However, the language and sentence construction used in abstracts is quite different from that used in the body of full text articles. There is now a sizable body of articles with licences that enable text mining (in 2020, about 600,000 full text, open access articles were deposited in Europe PMC, bringing the total available to 3.7M at the time of writing). In order to maximise the potential for reuse of this corpus it is imperative that text mining systems are trained appropriately, which means that there is a requirement for training sets/gold standards based on full text articles, not just abstracts. There are only two such training sets openly available to date: the Colorado Richly Annotated Full-Text Corpus (CRAFT)[16], which contains 97 full text open access biomedical journal articles, and more recently, NLM-Chem, which consists of 150 full text research articles annotated with chemical entities[17].

The CRAFT corpus has both semantic and syntactic annotations, including co-reference annotations and 10 biomedical concepts. The CRAFT corpus has become one of the most important gold standard datasets in the biomedical domain, having been mentioned in over 80 research papers. For example, OGER++[18] tool uses CRAFT corpus to train their feed forward neural network for disambiguation and normalisation (also referred as Concept Recognition) on top of the dictionary-based entity recognition system. In particular, evaluated on CRAFT corpus, more pronounced improvements have been achieved on Chemicals, Organisms, Proteins and sequences concepts. NLM-Chem is annotated with ~5000 unique chemical name annotations, mapped to ~2000 MeSH identifiers.

For over a decade, Europe PMC has routinely extracted a range of bioentities from incoming full text articles and abstracts, most recently applying this workflow to preprints. The text mining algorithm used is dictionary based, and is an extension of the Whatizit tool[5], used to extract:

1. Genes/Proteins
2. Organisms
3. Diseases
4. Chemicals
5. Gene ontology terms
6. Accession numbers (patterns)
7. Grants (patterns)
8. Resource names

The outputs of this pipeline are shared on the Europe PMC text mining community platform, via APIs and an annotation highlighting application called SciLite[19].

Recently we have been applying machine learning methodologies to improve the performance of this workflow and overcome the shortcomings of a dictionary-based approach, in particular issues such as false-positive identification of gene/protein names and false negative (non-identification) of disease states. In order to achieve this, we have developed a full text corpus training set: the Europe PMC Annotated Full Text Corpus (EPMCA). This corpus consists of 300 research articles selected from the Europe PMC Open Access subset annotated with three bio-entities: Gene/Proteins, Diseases and Organisms.

EPMCA represents the largest human annotated biomedical corpora to date, extending the contributions made by the CRAFT to the biomedical text mining community. This paper outlines the methodology used to generate the corpus, including article selection, dictionary tagging, and annotation guidelines.

The EPCMA corpus is open and free to use, and is available to download here: <https://gitlab.ebi.ac.uk/literature-services/public-projects/europepmc-corpus>

## Method

EPMCA was generated by human annotators in the triple-blind approach and the majority voting system was used for annotation acceptance. Annotators were asked to review bioentities pre-tagged by the existing Europe PMC text mining pipeline using the Hypothes.is web annotation tool, which is not only free to use but also uses the same underlying data structure as Scilite (the W3C recommended Web Annotations Data Model), allowing the human annotations to be integrated with existing text-mined annotations for further analysis. Furthermore, curation was based on annotator guidelines that were developed in agreement with the annotators. The guidelines proved to be useful for the annotators in selecting the correct entity type and text span. Figure 1 shows the full life cycle of the annotation process.

### Corpus assembly

As mentioned in the introduction, Europe PMC contains millions of full text articles and the manual annotation is a lengthy, complex and an expensive task. In order to create a representative training set, we needed to select a “typical” full-text corpus for manual annotation. Our first challenge therefore was to select the representative article set of 300 articles (number constrained by budget availability). We employed several techniques to stratify articles and select the representative set. We worked with Molecular Connections, India to employ three domain experts with PhD to annotate the corpus.

#### The Open Access article set in Europe PMC and CC-BY-licenced articles

As a primary outcome of this work was to create a training set for anyone to use, the first constraint applied was to use Open Access articles that have a parsable/machine-readable (available in JATS XML standard) CC-BY licence. We used the archived open access set from: 31st August 2018 (v.2018.09) [ Available at http://europepmc.org/ftp/archive/], which consists of 2,113,557 articles, of which 991,529 articles had a parsable CC-BY licence. Several additional rounds of selection were then applied to select a representative subset for the manual annotation.

#### Body Size

To find the most “typical” articles, we measured the size of the full text article <BODY> section and grouped them into bins of 10KB size. Most articles were in the range of 25-50 KB (Figure 2). Using this size range, we further constrained the pool to 503,950 articles. Advantage of constraining the article size meant that the annotators were provided with a more consistent article set in terms of length.

#### Entity frequency distribution

The pool of 503,950 “standard-sized” articles were further stratified based on the term frequency of three entity types of interest, namely gene/proteins, diseases and organisms. Using the Europe PMC dictionary-based annotation pipeline to annotate the articles, we established the range of entity frequencies in the articles (Figure 3) and created high (H), medium (M) and low (L) frequency tertiles by splitting them at the thirty three and the sixty six percentiles (Table 1). This resulted in 27 bins of articles from these tertiles of three entities ( 33 ) (Figure 4). The Low-Low-Low bin is sparsely populated with entities yet represents the largest number of articles. As these were considered to add little value to the training dataset, therefore it was excluded from the article selection process. We then randomly selected articles from each bin in proportion to the number of articles in each bin (2-20 articles from a bin in real terms, Figure 5).

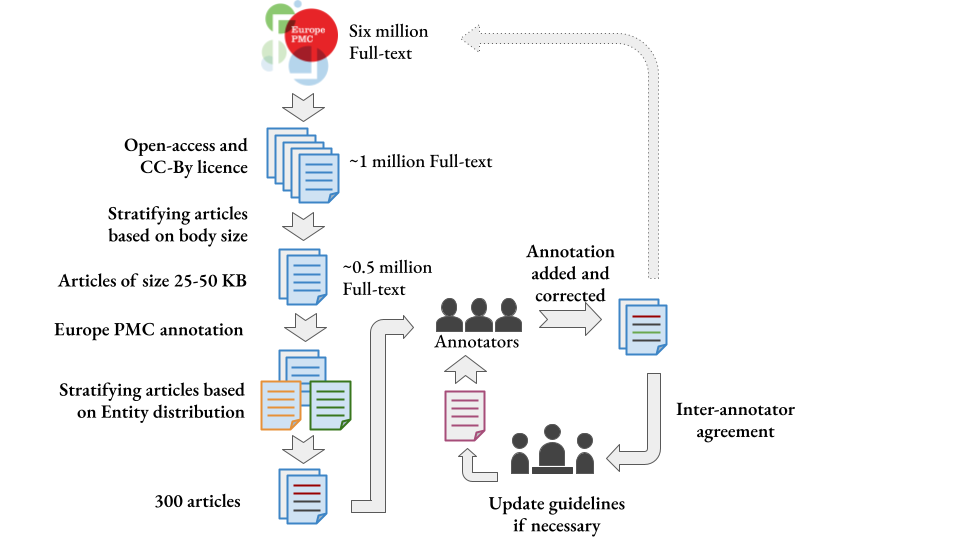


Figure 1: Full-text annotation life-cycle. Open Access subset of Europe PMC, archived on 31st of August, 2018 (v2018.09) contains about one million full-text articles with CC-By licence. We choose the gold standard articles from this subset using a number of stratification techniques. After analysing the article body size, the range of 25-50 KB was selected for further selection criteria. The entity frequency distribution from Europe PMC annotation pipeline was the next stratification criteria. The articles were categorised into Low, Medium, High bins for each entity type, i.e. Gene/Protein, Disease and Organisms. The articles from the Low-Low-Low (in order of Disease-Gene/Protein/Organism) were excluded from the final selection process due to their low value for the corpus. We worked with the annotators in synergy to improve the annotation guidelines.

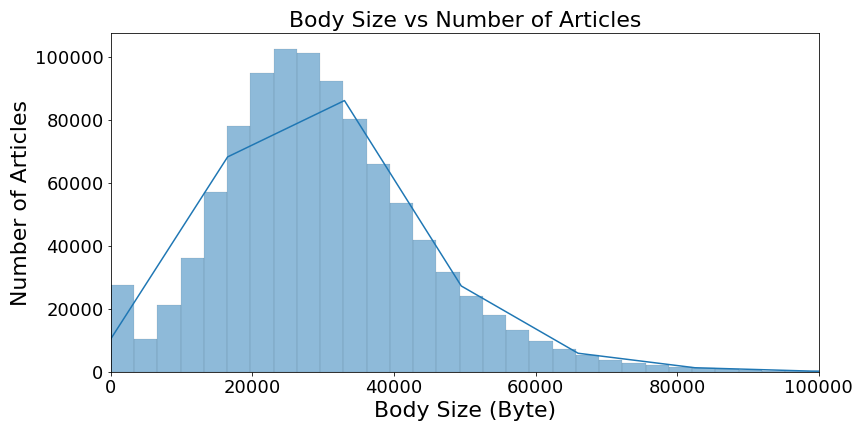


Figure 2: Distribution of body sizes of full text articles with a CC-BY licence in the 31st August 2018 (v.2018.09) frozen set.

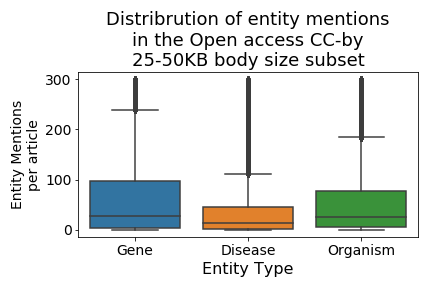


Figure 3 Distribution of entity, Gene, Disease and Organism, mentions per full-text article from the candidate pool. We have used a threshold of maximum 300 mentions per article per entity type for this figure. This figure shows on an average disease mentions are almost half or gene/protein mentions per article. This distribution helped us to set entity count boundaries for the article stratification and selecting the final corpus.

Table 1. Abundance of key entities used to establish tertile boundaries.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Entity | Low frequency of occurrence (L) | | Medium frequency of occurrence (M) | | High frequency of occurrence (H) | |
| Lower | Upper | Lower | Upper | Lower | Upper |
| Genes/Proteins | 0 | 11 | 12 | 80 | 81 | 2408 |
| Organisms | 0 | 9 | 10 | 57 | 58 | 3108 |
| Diseases | 0 | 4 | 5 | 32 | 33 | 678 |

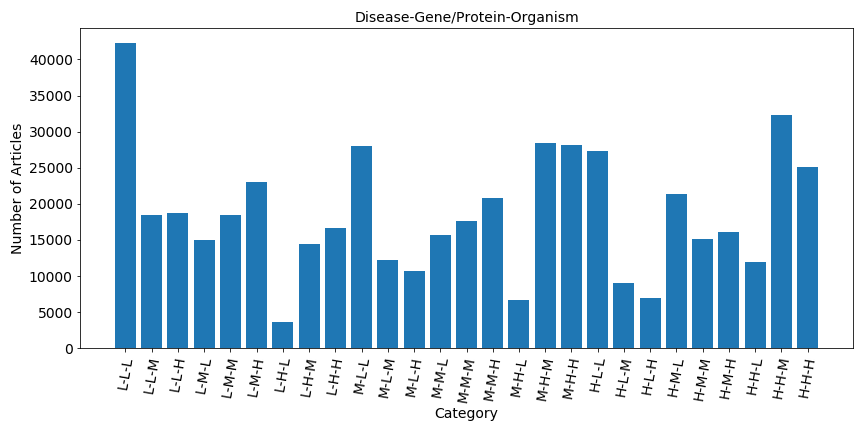


Figure 4. Distribution of articles based on the entity frequency. Here L, M and H represent Low frequency, Medium frequency and High frequency tertile. The order of the label is disease, gene/protein and organism. For example, H-L-H, represents articles that are high frequency for disease and organism and low frequency for gene/protein.

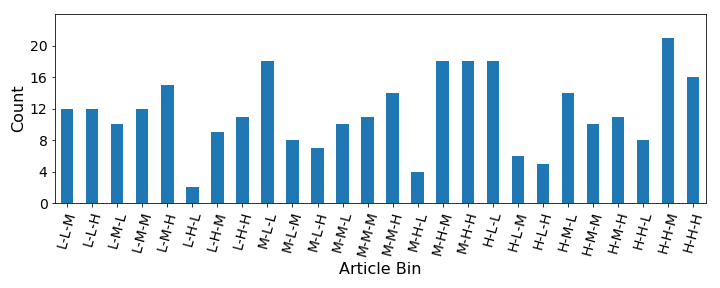


Figure 5: Proportional article distribution of gold standard corpus from the 26 bins.

### Ontology/terminology selection

The Europe PMC annotation pipeline uses a dictionary-based approach to tag gene/protein, disease and organisms[20]. The term dictionaries are created from UniProt, UMLS and NCBI taxonomy for the gene/protein, disease and organism respectively. The pipeline annotates articles using predefined patterns and regular expressions to accommodate term variations from the dictionaries. A list of common English words is used to avoid predominantly false positive identifications, for example, ‘CAN’ as a gene name. The annotators corrected these errors and removed other wrong annotations and added any missing terms. They were encouraged to use the same resources as reference for the new terms in order to normalise/link the terms to a standard entity database but were allowed to add novel terms using their domain expertise.

Gene/Protein

The gene/protein dictionary is periodically generated from the Swissprot data dump. Swissprot is a manually reviewed resource of proteins and genes. A gene/protein dictionary is generated from the gene name lines and their aliases. We used the UniProt version dated: 2014. We use a manually curated list of common English words, commonly used abbreviations of other entities along with the dictionary to reduce false-positive gene/protein identification.

Disease

UMLS disease terms are used to create the disease dictionary. In UMLS there are 12 different disease (DISO) groups, out of these 4 are used to generate the disease dictionary, as the other groups mainly comprise phenotypes and symptoms. The 4 DISO groups are as follows:

* T047 Disease or Syndrome
* T048 Mental or Behavioural Dysfunction
* T191 Neoplastic Process
* T046 Pathologic Function

ULMS version dated: 2015

Organism

The organism dictionary is based on the NCBI Taxonomy. Specific fields, such as acronym, blast name, genbank common name and genbank synonym are used to populate the dictionary.

NCBI taxonomy version dated:2015

### Creation of annotation guidelines

A detailed concept annotation guideline is essential to develop a good corpus and resolve annotation disputes (Supp data file [annotation guideline](https://docs.google.com/document/d/1mFiPy4yaRIA2Nix7OU-6S1wLxvhB3SJMnp6cY7dI20Q/edit#heading=h.gng9yl117ke)). CRAFT corpus provides comprehensive annotation guidelines[21] explaining the annotation text spans and the assignment of entity type. We based our annotation guideline on the CRAFT corpus guideline and expanded it to meet our requirements. A list of examples was included in the guidelines to assist curators. Before the start of the annotation work, a pilot study was conducted to annotate 3 articles. The aims of the pilot study were fourfold: First, the pilot study helped curators estimate the workload to set timelines for the project. Second, initial feedback was used to improve the annotation guidelines. Third, it gives curators the chance to get familiar with the task and annotation tools. Finally, it establishes the communication channel to manage the project.

### Article annotation

We worked with Molecular Connections, India to employ three domain experts with PhDs to annotate the corpus. We employed a triple-blind approach to annotation. Annotation discrepancies were resolved by the majority vote to achieve/ensure the best quality annotation, i.e., at least two annotators must agree on the annotation boundary and the entity type of the entity terms to pass the acceptance threshold. Moreover, it allows us to maximise the total number of annotations. For example, if a term is missed by one annotator, it is likely that it will be picked by the two other annotators. The triple-blind method made it possible to conveniently assess the inter-annotator agreements to ensure the quality of the annotation.

We sent the articles in multiple batches to the annotators to evaluate the annotation quality, inter-annotator agreement, address any confusion or quality issues and update the annotation guidelines if necessary after each batch. The 300 articles in the corpus were split into four batches. In order to assess the quality of the annotations, the first batch consisted of only 30 articles, after which the number of articles per batch increased. This approach allowed us to resolve annotation discrepancies along the way and refine the annotator guidelines. Table 2 shows a detailed breakdown of these batches.

Table 2: Batch wise annotation breakdown of articles and annotations.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Annotator1** | **Annotator2** | **Annotator3** | **Total** | **PMC count** |
| Batch1 | 1583 | 1587 | 1587 | 4757 | 30 |
| Batch2 | 4745 | 4727 | 4733 | 14205 | 70 |
| Batch3 | 5604 | 5610 | 5611 | 16825 | 80 |
| Batch4 | 11932 | 11924 | 11931 | 36787 | 180 |

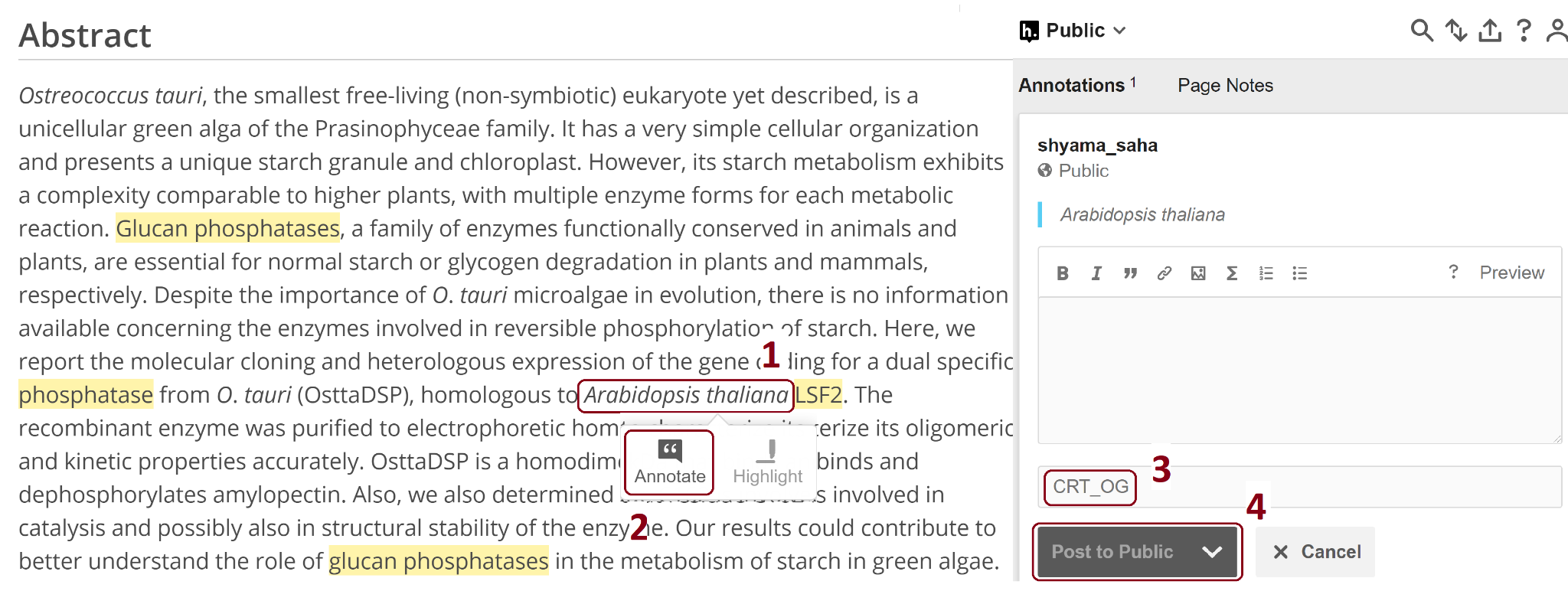


Figure 6: A screenshot of the hypothes.is annotation platform overlayed on top of the Europe PMC website. After selecting the term (1), users need to click the ‘Annotate’ button (2) to annotate the term. It will pop up the hypothes.is annotation window on the right-hand side, allowing the annotators to add the annotation (3) and then save it using the ‘Post to Public’ button (4). Please refer to the supplementary material (Section ‘How to Use the Interface’ in supplementary material [demo to molecular conenctions.docx](https://docs.google.com/document/u/1/d/1og34Ij2rf1zZPYgcA62sNK5YLOoYa0jp/edit?usp=drive_web&ouid=107119986075849487318&rtpof=true)) and hypothes.is website for detailed user manual.

Annotators were instructed to view the articles on the Europe PMC website, where they are displayed with the existing dictionary-based annotations from Europe PMC text-mining pipeline using Scilite. The hypothes.is annotation tool works as a layer on top of the Europe PMC website, allowing the curators/annotators to visualise and curate existing annotations and newly identified entity terms (Figure 6). We used Hypothes.is platform to eliminate the pre-processing steps required by the other annotation platforms such as BRAT[22] and GATE[23]. We developed a set of standard schemes of tags for the curators to use to classify the existing SciLite annotations.

The standard terms/tags were used as follows (Figure 7 shows an example of the use of these tags):

1. Correctness of annotation. Allows the annotators to verify existing Europe PMC annotations as Wrong Type (WT), Wrong Span (WS), Missing (MIS) or Correct (CRT).
2. Entity type. Three symbols representing the entity types, GP for Gene/protein, DS representing Disease and lastly organism by OG.
3. A special tag ‘ALL’. Allows the annotators to apply the annotation of the current term to all occurrences of it across the article. This was useful in the case of reducing workload for the annotators and annotation cost but required additional work to find all the occurrences of a concept with an “ALL'' tag in a post-processing phase.

These tags were used in combination to fully curate the annotations generated by the existing Europe PMC pipeline. For example:

* A correctly annotated gene/protein ( both entity type and annotation boundary), would be marked CRT\_GP.
* A wrong disease annotation would be marked WT\_DS, and if it should have been an organism would be marked: [WT\_DS][OG].

Finally, although entity annotation was the main goal of this initiative, annotators were asked to correct previously annotated gene/protein and disease associations. Therefore, the third tag scheme for the association annotation, YGD, NGD and AMB representing Correct relationship, Wrong relationship and ambiguous respectively.

A full list of all possible tags is in the section 1.1 [supplementary material](https://docs.google.com/document/d/1pyoI5sPwhX2bO3azJvvjA7MwV4EIxL9RcsMVZS2UiOE/edit).

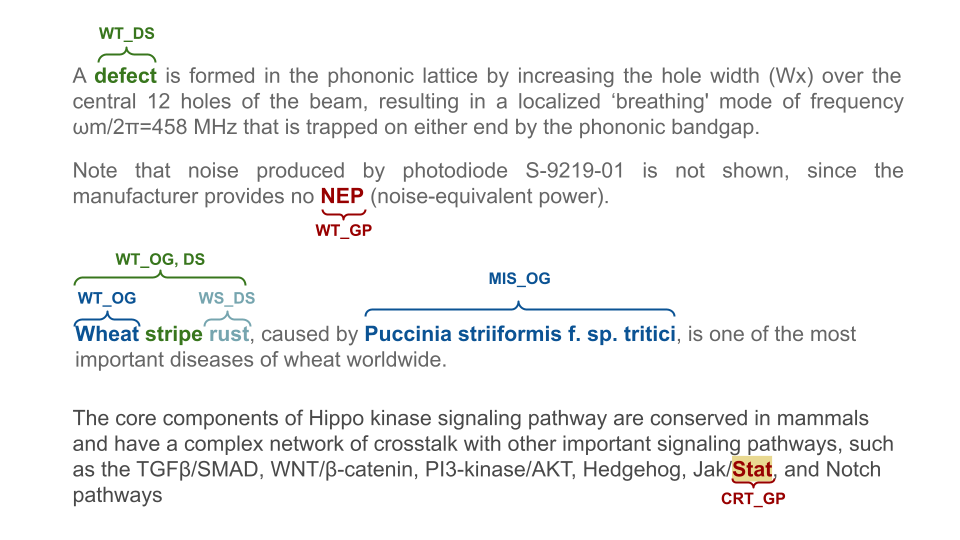


Figure 7: Example of manual annotation correcting dictionary-based Europe PMC annotation using the tag set defined for this annotation task. [describe the tags, or refer to the tag set description]

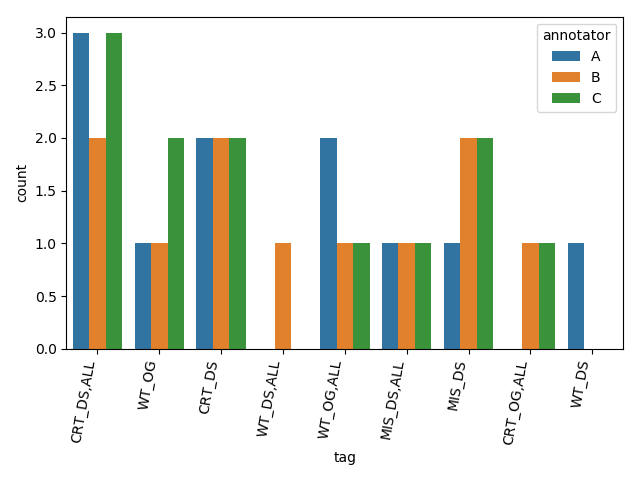
Figure 8 shows the use of these tag sets for assessing the annotation discrepancies among three annotators. 

Figure 8: An example of the tag distributions from batch 1 showing the discrepancies between the annotators. Annotators used the ‘ALL’ tag to mark all mentions of the entity as correct (CRT) or wrong type (WT), missing (MIS) and so on. The DS, OG represents the disease and the organism entities respectively.

### Annotation extraction and processing

Hypothes.is [https://web.hypothes.is/] is a free, open and user friendly platform enabling annotation of web content. The annotators used Hypothes.is to highlight the span of the entity terms, add notes and tag them with one of the available tags. They reviewed and marked pre-annotated terms as correct or incorrect and saved them using the hypothes.is platform.

At Europe PMC, sentence boundaries are added to the article XML files using an in-house sentenciser prior to entity recognition. The Europe PMC text-mining pipeline annotates the bio-entities using a dictionary-based approach and displays on the front-end HTML version via the web application called SciLite that requires further processing of the annotated XML file. The hypothes.is platform works on the front-end html version of the article. Each annotator set up an Hypothesis.is account and thus their annotations were saved to the Hypothes.is server (Please refer to Section ‘How to Use the Interface’ in supplementary material [demo to molecular conenctions.docx](https://docs.google.com/document/u/1/d/1og34Ij2rf1zZPYgcA62sNK5YLOoYa0jp/edit?usp=drive_web&ouid=107119986075849487318&rtpof=true) for detailed instructions). We retrieve the annotations using the Hypothe.is API in JSON format and it is converted to a CSV format using in-house tools. The hypothes.is JSON reports the annotated terms and their locations in respect to the HTML version of the article.

The annotations from the JSON file were extracted or tagged in the sentencised XML file using regular expressions. However, due to the inconsistency between the HTML article page and the XML file, a small number of annotations could not be successfully extracted using regular expressions. We have identified that the failure often occurs when an annotation is in a table. We post-processed the Hypothe.is JSON files for presenting the corpus for the wider community in multiple formats. More details in the following sections. Figure 9 shows an overview of the process.

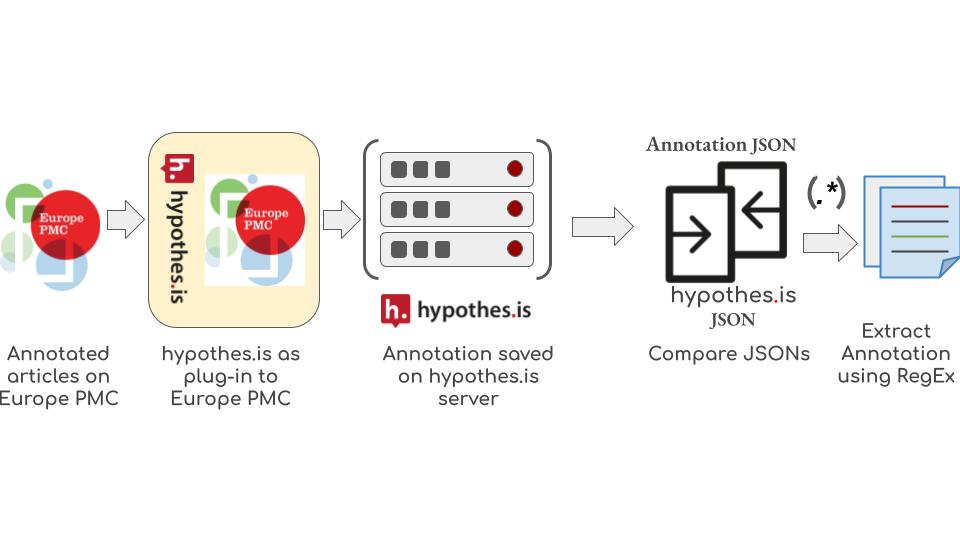


Figure 9: Annotation extraction workflow. hypothes.is was added onto Europe PMC as a plug-in for the annotation work. Annotators saved their annotations to the hypothes.is server in JSON format and it was retrieved and converted to CSV format using in-house tools. Europe PMC parses the XML version of the articles for sentence tagging and annotating named entities and displays a HTML version on the front-end. We compare the hypothe.is annotation JSON files against the XML version and extract the annotations using regular expressions.

## Results

In this paper, we present a corpus of three hundred full-text open-source articles from the biomedical domain, manually curated with the entities gene/protein, disease and organisms. Eight articles from the corpus do not contain any entity annotations, as all dictionary-based annotations were removed by the manual annotators. These articles came from 5 different bins. Table 3 (a, b) shows an overview of the manually annotated terms and compares these to the existing Europe PMC dictionary-based approach. For the purpose of evaluation of the dictionary-based approach, we applied majority voting acceptance criteria on the granular level annotation tags i.e. entity type tags (GP, DS, OG) along with the correctness tags (CRT, MIS, WT, WS).

Table 3 (a): This table shows the overall annotation statistics of the manual and the Europe PMC annotations for the selected three hundred gold standard articles. Overall, we have gained around 11k term annotations, with the highest gain coming from the gene/protein category. We report unique term count based on the string match and how many normalise to a database identifier rather than unique database identifier counts.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | Europe PMC Dictionary Based | | | | Gold Standard manual annotation | | | |
| Gene/Protein | Disease | Organism | Total | Gene /Protein | Disease | Organism | Total |
| Annotations | Total | 28,869 | 10,515 | 18,040 | 57,425 | 36,369 | 14,518 | 21,491 | 72,378 |
| Unique | 3,419 | 1,752 | 1,700 | 6,871 | 5,600 | 2,037 | 2,347 | 9,970 |
| Normalised to a DB entry | Total | - | - | - | - | 21,664 | 8,476 | 16,021 | 46,161 |
| Median per article | Total | 53.5 | 19.5 | 34 | 170 | 54.5 | 16 | 30 | 192 |
| Unique | 12 | 8 | 8 | 36 | 13 | 6.5 | 8 | 44.5 |
| Maximum annotation per article | Total | 722 | 219 | 407 | 955 | 795 | 478 | 456 | 940 |
| Unique | 113 | 78 | 111 | 156 | 178 | 76 | 170 | 201 |

Table 3(b): Evaluation of current Europe PMC dictionary workflow against the manual annotation. This table shows the number of dictionary based Europe PMC annotations updated by the manual annotators. A large proportion of the Europe PMC annotations are confirmed as correct by the manual annotators, although they also added/annotated a significant number of previously unidentified/un-annotated terms. A proportion of these terms are missed by the Europe PMC pipeline due to outdated dictionaries. The removed terms are often common English words or short acronyms. Gene/protein terms are more likely to be removed than other entity types ​​due to the frequency of occurrence and the false positive rate for three-letter GP acronyms. This row also counts the annotation where the dictionary based approach wrongly assigned the type, e.g. entities with the WT\_GP, DS tag will be added to the ‘removed’ cell count for the gene/protein and ‘added’ cell for the disease. The “Modified” row shows the number of entities that were modified/split into multiple entities. The overall column is the summation of correctness tags, i.e., CRT, MIS and WS going under the Correct, Added and the modified rows. For the WT tag, we are splitting them into two, one goes under the Removed column and the rest under the Modified row. When an annotation is assigned WT\_GP, it means that it's a wrong gene/protein annotation and removed from the annotation set. Whereas [WT\_GP, DS] tag means the annotation is not removed from the annotation set, but the entity type is modified.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Gene/Protein | | Disease | | Organism | | Overall |
|  | Unique | Total | Unique | Total | Unique | Total |
| Correct | 2,551 | 20,832 | 1,309 | 7,518 | 1,351 | 15,353 | 43,703 |
| Added | 2,671 | 13,718 | 575 | 5,836 | 820 | 5,307 | 24,230 |
| Removed | 697 | 6,172 | 447 | 1,991 | 207 | 982 | 8,514 |
| Modified | 561 | 1,819 | 269 | 1,164 | 311 | 831 | 4,445 |
| Precision | 0.72 | | 0.70 | | 0.89 | | 0.77 |
| Recall | 0.60 | | 0.56 | | 0.74 | | 0.64 |
| F1-score | 0.65 | | 0.62 | | 0.80 | | 0.70 |

The triple blind annotation approach had an overall inter-annotator agreement of 0.99. At this level we assigned granular tags to appropriate entity types, e.g., CRT\_GP, WS\_GP tags were mapped to GP tag, and used the MUCs’ strict evaluation rule for the inter annotator agreement. High inter-annotator agreement with the strictest methods shows that most of the annotations were agreed by all three annotators (Table 4). A total of 767 annotations are discarded since they were only annotated by one annotator. Among these discarded annotations, 289 annotations have overlapping text spans with the 1,005 annotations agreed by two annotators. For example, two annotators annotated “Welsh Mountain sheep”, however, the third annotator only annotated “sheep” from “Welsh Mountain sheep”. Both are correct in terms of definition of species. Only 478 annotations were truly discarded, accounting for 0.7% of total annotations. Further inspection of the discarded annotations may validate some and help keep the correct ones, but we did not consider this to be a major blocking task.

Table 4: Inter-annotator agreement statistics. We evaluated annotation agreement using SemEval-2013 Task9.1 strict rule. According to the strict evaluation rule, annotation agreement is reached only when two annotators agree on the term span and type of the annotation. We achieved an overall agreement of 0.99. First row of this table shows the entity level breakdown of annotations that were rejected due to the voting system, i.e., at least two annotators must agree on the annotation term, boundary and the entity type. Some of these entities were annotated by the other annotators with different entity boundaries.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Agreed by | Gene/Protein | Disease | Organism | Overall |
| 1 annotator | 270 | 178 | 319 | 767 |
| 2 annotators | 480 | 309 | 216 | 1005 |
| 3 annotators | 35934 | 14237 | 21298 | 71469 |

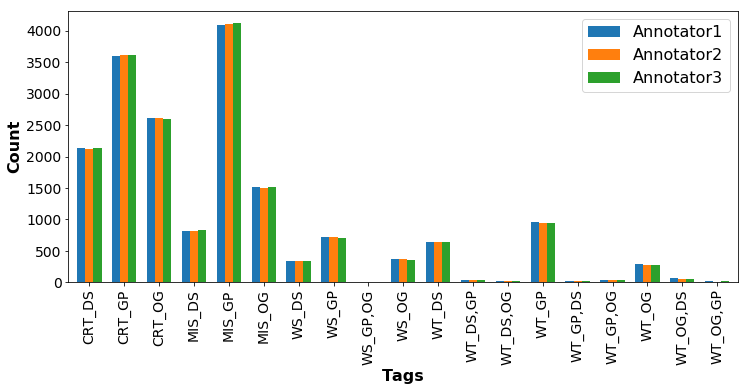


Figure 10: Entity tags distribution of the corpus and the comparison among the annotators. A large number of gene/protein terms are missed by the dictionary annotation. This figure demonstrates high inter-annotator agreement. Here the keywords CRT, MIS, WS and WT represent correct, missed, wrong span and wrong type respectively. The latter part of the tag represents the entity type namely, disease (DS), gene/protein (GP) and organisms (OG). Annotators use the WT keyword to remove an annotation and to change the entity type of an annotation. They submit the correct entity type by adding the correct entity type keyword after the WT tag, e.g., WT\_OG, DS.

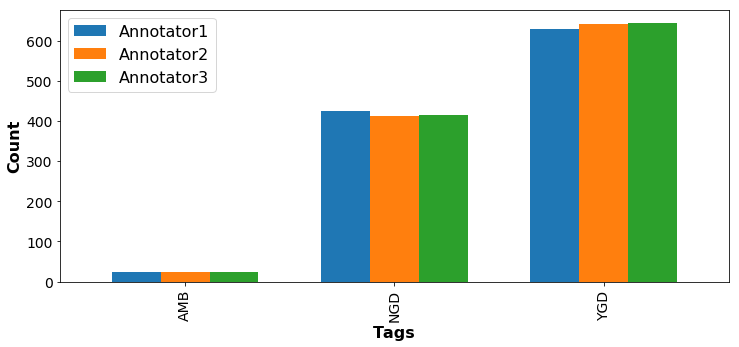


Figure 11: Although this annotation project mainly focused on the entity annotation, annotators were asked to annotate gene/protein-disease co-occurring sentences with entity relationship tag set. Please refer to the supplementary material section 1 (Tag schemes for annotation) for further details. This annotated set can be used to remove false positive association sentences and identify the challenging ones (AMB).

Our analysis on the distribution of tags set (Figure 10) shows the highest number of missing terms by the dictionary-based approach is from the gene/protein type (MIS\_GP tag). This might be since our gene/protein dictionary was last updated in 2014. Updating an entity dictionary involves several manual curations making it difficult to maintain. Although, a small number of these terms (1.6% of the terms tagged as MIS\_GP) were found in our gene/protein blacklist, showing the limitation of using the blacklist approach to restrict false positives. Using this gold standard data to train the state-of-the art machine learning/deep learning models for entity recognition eliminates these challenges. We observe the same trend for the false positive identifications, i.e., WT\_[GP|DS|OG]. The highest number of false positives are from the gene/protein type followed by the disease and organism terms respectively. The wrong type of annotation counts is quite low, that is, annotators only correct the entity type for a small number of annotations. This perhaps reflects the way the Europe PMC annotation pipeline works. This pipeline applies dictionaries sequentially, first the gene/protein dictionary, followed by the disease dictionary and then the organism dictionary. Once an entity is tagged, it becomes unavailable to tagging with subsequent dictionaries, which would likely reduce false positive disease and organism entity identifications. Our analysis shows only a few terms were assigned to the wrong entity type due to this approach proving our sequential method works. Table 5 shows how many terms annotations were updated to reassign the entity type.

Table 5: Europe PMC dictionary-based entity annotation follows a sequential manner to annotate the entities. For example, we apply the gene/protein dictionary before the disease dictionary making the gene/protein terms unavailable for the disease tagger. We minimise the false positive identifications through this approach. This table shows the number of wrong entity type assignments by the Europe PMC approach corrected by the manual annotators. Europe PMC misses a small percentage of the disease and organism entities due to the sequential approach. We are showing Europe PMC annotation in the rows and the manually corrected ones in the columns.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Manual annotation | | | |
| Europe PMC Annotation |  | Gene/Protein | Disease | Organism |
| Gene/Protein | - | 324 | 113 |
| Disease | 47 | - | 18 |
| Organism | 19 | 110 | - |

The special ‘ALL’ tag was used to indicate that annotation of a term applies to all occurrences of the term within the article. This was a significant time saver for articles that mention a particular entity ten or 100s of times. A total of 23,281 (7,336 unique) terms were tagged ‘ALL’.

Because Hypothes.is allows free text in the tag field, we identified a small number of errors in the tag names, e.g., ten annotations from annotators 1 and 2 uses ‘DIS’ instead of ‘DS’; one annotation uses ‘CRt’ instead of ‘CRT’. We corrected these errors for downstream analysis.

The titles of sections within a research article can vary widely but typically fall into a small number of categories. For example, “Methods” and “Methods and Reagents” are both classed as Methods sections. In the Europe PMC annotation pipeline, section titles are normalised to a set of 17 titles[24]. Figure 8 shows the entity distribution across these sections. As anticipated, it shows a high frequency of entity mentions in the main sections of an article, and furthermore demonstrates the value of full-text annotation vs using only abstracts. This entity distribution may help design a targeted annotation approach when resources are limited.

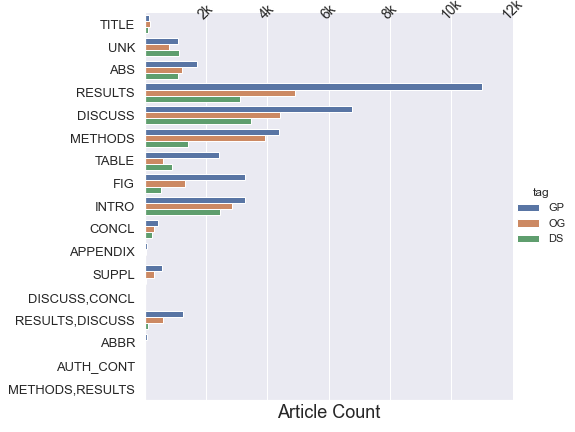


Figure 8: Term frequency distribution across different sections. Result, discussion, method and introduction sections contain the highest number of entity mentions. Gene/Proteins mentions in tables and figure titles are significantly higher than disease and organism mentions.

### Data availability

We have made this training set available in the following formats at <https://gitlab.ebi.ac.uk/literature-services/public-projects/europepmc-corpus>

1. Stand-alone curator annotations.
   1. CSV
   2. JSON
   3. IOB
2. Full text XML files (without EPMC annotations)
3. Full text XMLs with sentence boundary (we add <SENT> tag to annotate the sentence boundary)
4. Europe PMC annotation in JSON format.

To fit the diverse needs of the annotations, the corpus provides multiple formats of annotations from raw annotations of Hypothes.is platform (in csv format) to the standard and ready-to-use Inside-outside-beginning (IOB) format. In addition to the annotations, original full text articles are released in XML format without the tags. With the raw annotations in csv format and full text XML files, researchers are free to apply their own text mining tools to extract the annotations. The comma-separated values (CSV) raw annotation files contain three fields (exact, prefix and suffix) that are critical to locate the human annotations. “exact” is the annotation itself while “prefix” and “suffix” are characters before and after the annotation respectively. By combining “prefix”, “exact” and “suffix”, the snippet can locate the annotation using regular expressions. Raw annotations from all three human annotators are available, which are helpful for studies of agreement between annotators. Annotations in JavaScript Object Notation (JSON) and IOB formats are provided in addition to raw annotations. Both JSON and IOB format annotations are pre-processed in the way that only human annotations agreed by at least two annotations are included. The IOB format provides sentences with IOB tags and follows the CoNLL NER corpus standards. While the IOB format is widely used in named entity recognition (NER), researchers may prefer other tagging formats. As such, the JSON format provides sentences and annotations for researchers that are interested in transforming annotations into other tagging formats.

Moreover, full text articles are also available in the format that articles are split into sentences by the Europe PMC text mining pipeline. These sentences are used to extract human annotators, i.e., the annotations in the JSON and IOB formats.

## Discussion and Conclusion

Good quality gold standard datasets are essential to capitalise on the revolutionary deep learning methods for NLP and text-mining. While a number of gold standard datasets are available for biomedical entities, all except the CRAFT corpus and NLM-Chem are based on abstracts. As demonstrated here, high numbers of the entities appear in the body part of the articles. Moreover, the abstract is often a summary of the rest of the article and has very different language constructs. These differences make the gold standard abstract datasets unsuitable for full-text article annotation tasks.

The CRAFT corpus is a manually curated dataset comprising 97 full-text articles annotated with seven different biomedical ontologies. Despite the elaborate annotation scheme/entity types, it does not entirely overlap and cover the entity types we are interested in for the Europe PMC pipeline. This EPMCA corpus is larger than the CRAFT corpus and annotates gene/protein, disease and organism entity types and comprises ~1000 sentences with gene/protein and disease association annotations making it a unique addition for this domain. The approach to construction the annotation set is also slightly differently. We used triple-blind approach rather than the lead annotator approach for the CRAFT corpus and achieved a very high inter-annotator agreement across the corpus and all entity types. Use of the pre-annotated articles for the curation is another important difference between the CRAFT corpus and the Europe PMC corpus. We hope that the two approaches complement each other to support the future development of deep learning approaches

Pre-annotated terms and the use of the tag sets assisted the curators and allowed us to evaluate the dictionary-based annotation method in unprecedented detail. For example, the annotation tag ‘WT\_OG, GP’ allowed us to identify that ‘E9’ is a gene/protein term that is wrongly tagged as an organism by the existing annotation pipeline in the following sentence ‘Remodelling increases the off-rate a million-fold relative to that expected for a slip bond, allowing Im9 release and E9 activation at a biologically relevant rate upon binding to a competing organism’. We run the annotation pipeline sequentially in the following order: gene/protein, disease and organism and use blacklisted gene/protein terms (common English terms, disease name etc.) that removes those blacklisted terms to reduce false positive gene/protein terms. One of the challenges of dictionary-based annotation is manually curating and preparing the blacklist. Identifying ‘E9’ as an organism shows using blacklist can introduce false negatives and false positives. Without the context of the sentence and the paper, it is impossible to distinguish the type of the entity. Deep learning and machine learning methods will exactly help us address these challenges without the need of manual intervention using this corpus.

We are currently using the corpus to train Deep learning methods to address the challenges mentioned above for the named entity recognition (NER). Moreover, we have achieved a better sentence co-occurrence-based target-disease association identification through the improved NER system. Association annotation process is being further improved using the association annotation data set of the corpus. We hope this dataset will help the community in a similar fashion.

## Acknowledgements:

To Be Completed.

## Reference:

1. Samalikova, M. and R. Grandori, *Protein charge-state distributions in electrospray-ionization mass spectrometry do not appear to be limited by the surface tension of the solvent.* J Am Chem Soc, 2003. **125**(44): p. 13352-3.

2. Zahn-Zabal, M., et al., *The neXtProt knowledgebase in 2020: data, tools and usability improvements.* Nucleic Acids Res, 2020. **48**(D1): p. D328-D334.

3. Wei, C.H., et al., *PubTator central: automated concept annotation for biomedical full text articles.* Nucleic Acids Res, 2019. **47**(W1): p. W587-W593.

4. Cheng, D., et al., *PolySearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites.* Nucleic Acids Res, 2008. **36**(Web Server issue): p. W399-405.

5. Rebholz-Schuhmann, D., et al., *Text processing through Web services: calling Whatizit.* Bioinformatics, 2008. **24**(2): p. 296-8.

6. Gerner, M., G. Nenadic, and C.M. Bergman, *LINNAEUS: a species name identification system for biomedical literature.* BMC Bioinformatics, 2010. **11**: p. 85.

7. Cook, H.V., E. Pafilis, and L.J. Jensen. *A dictionary-and rule-based system for identification of bacteria and habitats in text*. in *Proceedings of the 4th BioNLP shared task workshop*. 2016.

8. Li, J., et al., *BioCreative V CDR task corpus: a resource for chemical disease relation extraction.* Database, 2022. **2016**.

9. Kafkas, et al. *CALBC: Releasing the Final Corpora*. in *LREC*. 2012.

10. Kim, J.D., et al., *GENIA corpus--semantically annotated corpus for bio-textmining.* Bioinformatics, 2003. **19 Suppl 1**: p. i180-2.

11. Collier, N. and J.-D. Kim. *Introduction to the Bio-entity Recognition Task at JNLPBA*. 2004. Geneva, Switzerland: COLING.

12. Franzén, K., et al., *Protein names and how to find them.* Int J Med Inform, 2002. **67**(1-3): p. 49-61.

13. Doğan, R.I., R. Leaman, and Z. Lu, *NCBI disease corpus: a resource for disease name recognition and concept normalization.* J Biomed Inform, 2014. **47**: p. 1-10.

14. Sunil Mohan, D.L., *MedMentions: A Large Biomedical Corpus Annotated with UMLS Concepts.* arXiv, 2019(1902.09476).

15. Krallinger, M., et al., *The CHEMDNER corpus of chemicals and drugs and its annotation principles.* J Cheminform, 2015. **7**(Suppl 1 Text mining for chemistry and the CHEMDNER track): p. S2.

16. Bada, M., et al., *Concept annotation in the CRAFT corpus.* BMC Bioinformatics, 2012. **13**: p. 161.

17. Islamaj, R., et al., *NLM-Chem, a new resource for chemical entity recognition in PubMed full text literature.* Sci Data, 2021. **8**(1): p. 91.

18. Furrer, L., et al., *OGER++: hybrid multi-type entity recognition.* J Cheminform, 2019. **11**(1): p. 7.

19. Venkatesan, A., et al., *SciLite: a platform for displaying text-mined annotations as a means to link research articles with biological data.* Wellcome Open Res, 2016. **1**: p. 25.

20. Kafkas, Ş., J.H. Kim, and J.R. McEntyre, *Database citation in full text biomedical articles.* PLoS One, 2013. **8**(5): p. e63184.

21. Bada, M., et al. *An Overview of the CRAFT Concept Annotation Guidelines*. 2010. Uppsala, Sweden: Association for Computational Linguistics.

22. Pontus Stenetorp, S.P., Goran Topić, Tomoko Ohta, Sophia Ananiadou, Jun'ichi Tsujii *brat: a Web-based Tool for NLP-Assisted Text Annotation.* Organization. **1**: p. T2.

23. Cunningham, D.M.H. and K. Bontcheva, *Text Processing with GATE (Version 6)*. 2011: University of Sheffield D.

24. Kafkas, Ş., et al., *Section level search functionality in Europe PMC.* J Biomed Semantics, 2015. **6**: p. 7.