Files and annotation guidelines

Abstract

The Bio-ID track focuses on entity tagging and ID assignment to selected bioentity types, with the aim of facilitating downstream article curation both at the pre- and post-publication stages. This document describes details of the data set file format and of key annotation aspects of the corpus of figure panel legends provided by SourceData for the Bio-ID track. The task is to annotate text from figure legends with the entity types and IDs for taxon (organism), gene, protein, miRNA, small molecules, cellular components, cell types and cell lines, tissues and organs. There will be two subtasks: batch annotation of entities and IDs; and interactive annotation for entities and IDs, based on annotations provided from the batch subtask. Training materials will consist of SourceData annotated figure legends (by panel), in BioC format, and the corresponding full text articles (also BioC format) provided for context. Participants can participate by submitting annotations for one or more bioentity types, in BioC format.

We also plan to use BeCalm, the annotation metaserver (http://www.becalm.eu/) with available entity recognition services to process test corpora.

Table of Contents

SourceData

SourceData (http://sourcedata.embo.org) is a platform that allows researchers and publishers to share scientific figures and, when available, the underlying source data, in a way that is machine-readable and findable. SourceData focuses on the core of scientific evidence - data presented in figures - and makes papers searchable based on their data content; for more information, see (http://biorxiv.org/content/early/2016/06/20/058529).

Task

The Bio-ID track will focus on entity tagging and ID assignment for selected classes of bioentities, with the aim of facilitating downstream article curation at both the pre- and post-publication stages. To further this, we are bringing together multiple stakeholders to discuss functional requirements and develop interoperable digital curation tools. This track builds on the SourceData project as well as previous BioCreative experiments, including the interactive tracks (IAT), the BioC and gene/protein/chemical extraction tracks, and the BeCalm framework. The track is designed to foster the development of an integrated and interoperable workflow of multiple text mining tools for real-world testing in pilot publishing frameworks.

More information about this track can be found under Tasks or at
http://www.biocreative.org/tasks/biocreative-vi/track-1/

Bio-ID Training Data Set

The data set consists of 13,573 annotated figure panel captions corresponding to 570 full length articles. Note that a figure panel caption may include the text from more than one panel (e.g., PMC3791395, Figure_6-A-C).

Data Set Content

The training data set is in the caption_bioc folder which contains a set of files in BioC format. Each file consists of a collection of figure captions for a given article annotated by SourceData curators. The name of the file corresponds to its PMC ID and the corresponding full-text article is provided separately (in fulltext_bioc folder). Within a BioC file, each <document> element contains the annotation for each individual figure panel within the article; the <id> given is made up of the PMCID followed by the Figure_number-panel (e.g.,<id>5048346 Figure_1-A</id>).

Example of SourceData Annotated Data:

```
<document>
   <id>5048346 Figure_1-A</id>
   <infon key="sourcedata_document">2225</infon>
   <infon key="doi">10.15252/embj.201694885</infon>
   <infon key="pmc_id">5048346</infon>
   <infon key="figure">Figure 1-A</infon>
   <infon key="sourcedata_figure_dir">Figure_1-A</infon>
   <passage>
     <offset>0</offset>
     <text>A. The localization of NSUN3 was analysed in HEK293 cells stably expressing NSUN3-GFP
(green). NSUN3-GFP and staining with a Mitotracker (red) are shown separately and in an overlay with
DAPI to indicate nuclei. The scale bar represents 5 m.</text>
     <annotation id="1">
       <infon key="type">Uniprot:Q9H649</infon>
       <infon key="sourcedata_figure_annot_id">1</infon>
       <infon key="sourcedata_article_annot_id">1</infon>
       <location offset="23" length="5"/>
       <text>NSUN3</text>
     </annotation>
```

Within a <document> element, under <passage> the <*text*> element contains the caption text passage relevant to the figure panel. This is the text snippet that must be processed, to extract the relevant bioentity types along with their IDs from the appropriate standard resources (see Table 1 below).

Example:

```
<text>A. The localization of NSUN3 was analysed in HEK293 cells stably expressing NSUN3-GFP
(green). NSUN3-GFP and staining with a Mitotracker (red) are shown separately and in an overlay with
DAPI to indicate nuclei. The scale bar represents 5 m.</text>
```

Note that the text for the panel may consist of discontinuous text derived from the Figure caption; it may also include legends from several panels.

The <*annotation*> element contains the type, the id for annotation, the location of the annotation in the text passage (with location specified by offset and length, both in bytes), and the text annotated.

Example of SourceData Annotated Data:

```
<annotation id="1">
  <infon key="type">Uniprot:Q9H649</infon>
  <infon key="sourcedata_figure_annot_id">1</infon>
  <infon key="sourcedata_article_annot_id">1</infon>
  <location offset="23" length="5"/>
  <text>NSUN3</text>
</annotation>
```

Annotation types:

There are multiple annotation types annotated by SourceData in the corpus, but not all will be considered for the evaluation. Those to be considered are listed in Table 1.

Table 1. Entity types with corresponding resources that can be linked.

Entity Type	Resources	Example **infon key=”**type”	Generic infon **key=”type”**
Protein	UniProt	Uniprot:Q9H649	Uniprot:<Accession>
Gene	NCBI gene	NCBI gene:4137	NCBI gene:<ID>
miRNA	Rfam	Rfam:RF00076	Rfam:<ID>
Small molecules	ChEBI (primary)	CHEBI:15996	CHEBI:<ID>
	PubChem (secondary)	PubChem:5717066	PubChem:<id>
Cellular component	GO cellular component	GO:0005886	<GOID>
Cell types and cell lines	Cellosaurus (primary)	CVCL_U985	<accession>
	Cell Ontology (secondary)	CL:0000540	<CLID>
Tissues & organs	Uberon	Uberon:UBERON:0002048	Uberon:<UBERONID>
Organisms & species	NCBI Taxonomy	NCBI taxon:10090	NCBI taxon:<ID>

Sources: All ID sources are publicly available
UniProt: www.uniprot.org
NCBI gene: https://www.ncbi.nlm.nih.gov/gene
Rfam: http://rfam.xfam.org/
ChEBI: https://www.ebi.ac.uk/chebi/
PubChem: https://pubchem.ncbi.nlm.nih.gov/
Gene ontology: http://geneontology.org/

Cellosaurus: http://web.expasy.org/cellosaurus/
Cell Ontology: http://obofoundry.org/ontology/cl.html
Uberon: http://uberon.github.io/
NCBI Taxonomy: https://www.ncbi.nlm.nih.gov/taxonomy

What has been annotated in the SourceData corpus?

Here are some key considerations of the curation practice (extracted from the SourceData definitions for curators by Thomas Lemberger):

- SourceData description of the data presented in scientific figures specifies the entities that are relevant to the scientific meaning of the data. To be tagged by curators, a panel must report experimental data (relevant panel legend). Panels that present schematics, computational simulation results, overviews, workflows are not tagged.
- In the text of a relevant panel legend, terms that correspond to entities types on Table 1 are all tagged.
- Entities are assigned to only one entity type of those listed in Table 1.
- In general, generic terms referring to broad classes of biological components (e.g. 'proteins', 'cells', 'animals') are not be tagged unless they refer to the object of an assay.
- Some terms such as those referring to proteins or genes can be appended with prefixes or suffixes that indicate a post-translational modification, a mutation or other variations of the actual base term. In such case, pre- or suffixes are left out and only the base term is tagged (e.g., in text describing a mutant of B-RAF 'B-RAF(V600E)', only B-RAF is tagged; similarly with p-AKT1 that designates the phosphorylated form of AKT1, only AKT1 is tagged).
- In other cases, a prefix is added to an entity to denote a species origin, in which case the prefix should be kept (e.g., dMyc to denote the homolog in Drosophila of Myc)
- Some components are engineered by assembling or fusing multiple sub-components, these are tagged individually. For example, the term 'RAS-GFP' referring to a fusion protein between GFP and RAS is annotated with two tags: 'RAS' and 'GFP'.

Special considerations for the BIO-ID track

- Participating teams can provide annotation for one or more types of those described in Table 1.
- For this track, the gene/protein type can be treated interchangeably (i.e., all proteins and gene mentions can be linked to NCBI gene and/or UniProt identifiers).
- Some entity types, like small molecules, have more than type of possible identifier; however, one of the resources is considered primary. To help with training and comparison, we provide mappings, to the extent possible, between identifiers of the same entity type. However, the best practice would be to look up the entity in the preferred resource, and if it cannot be found, then look it up in the secondary resource.
- In particular, for UniProtKB, UniProtKB/Swiss-Prot (reviewed) entries should be linked whenever possible; TrEMBL (unreviewed) entries should only be linked when no corresponding Swiss-Prot entry is available.

- We will only compare the captions for entity types in Table 1 that SourceData has annotated and which have associated IDs. Both correct mention detection and ID normalization will be evaluated.
- Note that there are a number of entities annotated without identifiers – for scoring purposes we will ignore these.

Scoring

We plan to release a scorer for use in training. The scorer will provide a break down by object type for results compared against the SourceData annotated data. The scorer will provide details on mention level tagging as well as ID matching.

Description of the data set

570 papers annotated for figure captions from some 22 journals;
94,266 annotations with IDs (see table below);
110,416 annotations (including annotations without IDs – see annotations.csv for a listing)

Entity Type	Number of articles with given entity	Total number of given entity type	Number of unique given entity type
Protein/Gene (561 articles)	UniProt (548)	30211	2833
	NCBI gene (537)	21766	2451
miRNA	9	167	13
Small molecules (513 articles)	ChEBI (506)	9869	786
	PubChem (134)	700	140
Cellular component	482	7310	376
Cell types and cell lines (482 articles)	Cellosaurus (351)	5783	230
	Cell Ontology (300)	4638	217
Tissues & organs	316	5870	459
Organisms & species	454	7952	147

Note that this dataset is enriched in articles related to autophagy.

Additional files

We also are providing the following files:

- Full length articles for the documents in training set in BioC format (in the fulltext_bioc directory);
- The caption files without annotations (caption_bioc_unannotated) – this will be the format for the test data;
- A file with examples of input, output and SourceData annotations (input_output.txt)
- Reciprocal mapping between UniProt and NCBI gene, and between ChEBI and Pubchem (in the mapping_training.xlsx file);
- A listing of all the annotations in the 570 paper training set (annotations.csv)