PubAnnotation - a storage system for sharing of literature annotation

Jin-Dong Kim
Database Center for Life Science
jdkim@dbcls.rois.ac.jp

PubAnnotation is a storage system for sharing of literature annotation. It maintains texts from PubMed and PubMed Central (the open access subset) in a canonicalized form. Annotations, which are even produced without any connection to PubAnnotation, can be uploaded to the storage. What PubAnnotation does during the upload is to align the annotations, e.g., the character offsets, to the canonicalized texts, so that users do not need to worry about frequent small variations, e.g., extra spacing or different encoding of Greek letters, in the texts, which often causes a problem when compiling annotations from different sources. As a result, all the annotations uploaded to PubAnnotation become directly comparable to each other even if they have come from different sources or projects.

The annotations stored in PubAnnotation also can be downloaded to the local storage of a user. At that time, the user can specify a specific version of the texts of his/her own, to which the annotations to be downloaded will be aligned. In the way, the annotations become portable to variants of the base texts. It can be illustrated as follows:

$text\_v_1$ ← annotations → $text\_v_c$

All the annotations stored in PubAnnotation are aligned to the canonicalized version of the base text (indicated by (v_c)). A user may have the same texts but with some variations (indicated by (v_1)). Annotations made to either version of the text become portable to the different versions of text through PubAnnotation.

We experimented the functionality of PubAnnotation with three open corpora with annotations: Genia, AIMed, and Genetag. The base texts of those corpora were collected and preprocessed by the corpus developers using different pipelines. However, we found that all the annotations in the corpora could be successfully aligned to the canonicalized texts, which were taken from PubMed, by simply uploading them to PubAnnotation.

We expect the aligning technology implemented in PubAnnotation to substantially reduce the cost of the community to seek interoperability of literature annotation.