$BioC_{convert}$: A Conversion Tool Between BioC and PubAnnotation

Donald C. Comeau, Rezarta Islamaj Doğan, Sun Kim, Chih-Hsuan Wei, W. John Wilbur and Zhiyong Lu
National Center for Biotechnology Information
National Library of Medicine, NIH
Bethesda, MD 20894, USA
comeau@ncbi.nlm.nih.gov

***Abstract*— BioC is a simple XML data format for text, annotations, and relations. PubAnnotation is a repository of text annotations focused on the life science literature. A conversion tool between BioC XML and the JSON import / export format of PubAnnotation has been developed, BioCconvert. As a demonstration, the Ab3P gold standard abbreviation annotations are being made available through PubAnnotation.**

Keywords—BioC, PubAnnotation, interoperability, biomedical annotations

I. Introduction

BioC is a simple data structure for text, annotations, and relations [1]. It was developed to support the BioCreative series of workshops. It was successfully used in dedicated BioC tracks at BioCreative IV [2] and BioCreative V [3]. It was also used in other tracks such as the Comparative Toxicogenomics Database (CTD) Curation track at BioCreative IV [4] and the Chemical Disease Relation (CDR) track at BioCreative V [5]. BioC annotations are specific identified and labeled substrings of the original text. They do not need to be continuous. They occur in a passage, or sentence, along with, or parallel to the original text. Relations connect an arbitrary number of annotations, or other relations, in anyway desired. The details of a relationship should be described in an accompanying key file.

PubAnnotation is a repository of text annotations mainly developed and maintained by DBCLS (Database Center for Life Science) [6]. It focuses on annotations to the life science literature, particularly PubMed® abstracts and PubMed Central® (PMC®) full text articles. PubAnnotation allows for three types of annotations: denotations, relations, and modifications. A denotation is an indentified and labeled portion of the original text. This is what, in other contexts, is often simply called an annotation. A relation describes the relationship between two denotations, as expected. A modification changes a single denotation or relation. Supported examples are Speculation and Negation.

Both BioC and PubAnnotation have sizeable and growing communities. According to Google Schoolar, the original BioC paper has 60 citations. More than 15 papers on or using BioC appear in PubMed. The original PubAnnotation project has 8 citations. The PubAnnotation site lists 138 projects, of which 26 have been released. PubAnnotation corpora it would be nice to see in BioC include CoMAGC, a cancer and gene corpus, and SPECIES800, an organism corpus. BioC corpora that might be useful in PubAnnotation include DDIcorpus and GeneTag. BioC tools that could be applied to PubAnnotation corpora include abbreviation finding, NLP pipelines in C++ and Java and a number of NER tools. The benefits of interoperability between BioC and PubAnnotation are clear.

This research was supported by the Intramural Research Program of the NIH, National Library of Medicine.

II. Conversion and Example

PubAnnotation has a mechanism to add documents in addition to their existing PubMed and PMC sets. Since our example used PubMed references, no additional PubAnnotation documents needed to be created and this feature of PubAnnotation is not addressed. Only the appropriate annotations needed to be created or interpreted. When a PubAnnotation denotation is created, the text of the enclosing passage is reported. Modifiers are used to represent unary BioC relations, while relations represent binary BioC relations respectively. Offsets were adjusted to refer to the reported text. Lengths were used to calculate the end of a span. Table 1 shows sample BioC XML annotations and the corresponding PubAnnotation JSON.

The conversion tool ($BioC_{convert}$) is implemented in Python. In addition to having a BioC implementation, Python ships with a standard JSON library. As a demonstration of this tool, the abbreviation definition corpus created to test the Ab3P abbreviation definition identifier [7,8] was added to PubAnnotation. This gold standard corpus includes 1250 manually annotated MEDLINE records. It includes 1221 abbreviation-definition pairs. For an abbreviation definition, both the abbreviation (short form) and its definition (long form) are identified. There are a number of reasons this corpus was chosen as the demo corpus. The concepts of abbreviation definition is very simple and clear, so reviewing the imported annotations for accuracy was easy. Since the relationship between an abbreviation and its defining long form is explicict in the corpus, importing relations could be tested in addtion to just importing denotations.

Importing the corpus into PubAnnotation was tested in two ways. First, the imported corpus was exported in the PubAnnotation format and converted back to BioC. This stable round-trip precludes a large number of bugs. However, because the PubAnnotation format lacks redundancy, this roundtrip does not guarantee accuracy. The developers used visual tools

to manually review articles. This ensured the annotations were imported accurately. At this time, PubAnnotation does not support multi-segment denotations. Thirteen articles include at least one multi-segment abbreviation. These were given a span that covers all the individual spans. The Ab3P corpus is available at http://pubannotation.org/projects/Ab3P-abbreviations. BioC$_{convert}$ will be available via a link at http://bioc.sourceforge.net.

III. Discussion

BioC is a desirable datasharing format because while being a minimalistic approach, it is also very flexible, allowing a wide range of annotations to be represented. However, not everything in BioC can be represented in PubAnnotation. PubAnnotation allows for unary relations (modification) and binary relations (relation), while BioC allows for n-ary relations. However, unary and binary are by far the most common. If other relation types become more common, it is likely that PubAnnotation will support them.

BioC infons (key-value pairs) allow arbitrary additional information about each annotation to be recorded. Unfortunately, information beyond the object type will be lost in PubAnnotation. Nonetheless, the annotation will still be useful in the PubAnnotation repository. Since BioC allows arbitrary role labels in relations, manual configuration is required to ensure that the correct BioC information is recorded in the PubAnnotation relation "subj," "pred," and "obj" fields.

While the intent of BioC$_{convert}$ is to be general purpose, since it has been tested on only one corpus, it is likely task specific in unintended and undetected manners. Porting additional annotation collections between BioC and PubAnnotation will identify and allow correcting these deficiencies, if they exist.

IV. Conclusion

With the creation of BioC$_{convert}$, one can now convert between BioC XML and PubAnnotation JSON. It is possible for BioC tools to be applied to any of the annotations available from PubAnnotation. Conversely, annotations available in BioC can be shared via PubAnnotations.

References

[1] Comeau, D. C., Islamaj Dogan, R., Ciccarese, P., Cohen, K. B., Krallinger, M., Leitner, F., . . . Wilbur, W. J. BioC: a minimalist approach to interoperability for biomedical text processing. Database (Oxford), 2013, bat064. doi:10.1093/database/bat064.

[2] Comeau, D. C., Batista-Navarro, R. T., Dai, H. J., Dogan, R. I., Yepes, A. J., Khare, R., . . . Wilbur, W. J. BioC interoperability track overview. Database (Oxford), 2014. doi:10.1093/database/bau053.

[3] Kim, S., Islamaj Doğan, R., Chatr-aryamontri, A., Tyers, M., Wilbur, W. J., & Comeau, D. C. Overview of BioCreative V BioC Track. Paper presented at the Fifth BioCreative Challenge Evaluation Workshop, Seville, Spain, 2015.

[4] Wiegers, T. C., Davis, A. P., & Mattingly, C. J. Web services-based text-mining demonstrates broad impacts for interoperability and process simplification. Database (Oxford), 2014. doi:10.1093/database/bau050.

[5] Wei, C.-H., Peng, Y., Leaman, R., Davis, A. P., Mattingly, C. J., Li, J., . . . Lu, Z. Overview of the BioCreative V Chemical Disease Relation (CDR) Task. Paper presented at the Fifth BioCreative Challenge Evaluation Workshop, Seville, Spain, 2015.

[6] Kim, J.-D., & Wang, Y. PubAnnotation: a persistent and sharable corpus and annotation repository. Paper presented at the Proceedings of the 2012 Workshop on Biomedical Natural Language Processing, Montreal, Canada, 2012.

[7] Sohn, S., Comeau, D. C., Kim, W., & Wilbur, W. J. Abbreviation definition identification based on automatic precision estimates. BMC Bioinformatics, 9, 402. doi:10.1186/1471-2105-9-402, 2008.

[8] Islamaj Doğan, R., Comeau, D. C., Yeganova, L., & Wilbur, W. J. Finding abbreviations in biomedical literature: three BioC-compatible modules and four BioC-formatted corpora. Database: The Journal of Biological Databases and Curation, 2014, bau044. http://doi.org/10.1093/database/bau044

TABLE I. Equivalent BioC XML and PubAnnotation JSON for the same text and abbreviation definition annotations. The color coded sections indicate equivalent information. Yellow: text, Red: id, Purple: offset, Green: length, or end, of annotation.

BioC XML

```
<passage>
 <infon key="type">title</infon>
 <offset>0</offset>
 <text>Comparison of two timed artificial insemination (TAI)
protocols for management of first insemination postpartum.</text>
 <annotation id="SF0">
  <infon key="ABBR">ShortForm</infon>
  <infon key="type">ABBR</infon>
  <location offset="49" length="3"/>
  <text>TAI</text>
 </annotation>
 <annotation id="LF0">
  <infon key="ABBR">LongForm</infon>
  <infon key="type">ABBR</infon>
  <location offset="18" length="29"/>
  <text>timed artificial insemination</text>
 </annotation>
 <relation id="R0">
  <infon key="type">ABBR</infon>
  <node refid="LF0" role="LongForm"/>
  <node refid="SF0" role="ShortForm"/>
 </relation>
</passage>
```

PubAnnotation JSON

```
[
 { "denotations": [
    { "span": { "begin": 49, "end": 52 },
      "obj": "ABBR",
      "id": "SF0" },
    { "span": { "begin": 18, "end": 47 },
      "obj": "ABBR",
      "id": "LF0" }
  ],
  "target": "http://pubannotation.org/docs/sourcedb/
PubMed/sourceid/12018411",
  "sourceid": "12018411",
  "sourcedb": "PubMed",
  "relations": [
   { "pred": "ShortForm",
     "obj": "SF0",
     "subj": "LF0",
     "id": "R0" }
  ],
  "project": "Ab3P_abbreviations",
  "text": "Comparison of two timed artificial insemination (TAI)
protocols for management of first insemination postpartum." }
]
```