MEETING PROPOSAL

Towards seamless format conversion between BioC and PubAnnotation for sharing PubMed/PubMed Central documents and annotations

Sun Kim[1], Donald C. Comeau[1], Rezarta I. Doğan[1] and Zhiyong Lu[1*]

*Correspondence:
zhiyong.lu@nih.gov
[1]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

BioC [1] is a simple XML-based format designed to provide interoperability for text mining tools and manual curation results. Recently, BioC organizers redefined BioC as a data structure and XML/JSON as the means to transmit BioC data, i.e. use either XML or JSON for serialization. This is an important addition because JSON is the most common format for data interchange, rapidly replacing XML [2]. In addition, it is important to discuss and implement an integrated data exchange framework among standard formats such as BioC and PubAnnotation.

PubAnnotation is a repository of text annotations developed and maintained by DBCLS (Database Center for Life Science) [3]. Although it focuses on annotations to the life science literature, e.g. PubMed® abstracts and PubMed Central® (PMC®) full text articles, PubAnnotation also defines a JSON annotation format [4] for communication. Comeau et al. [5] implemented a conversion tool between BioC XML and PubAnnotation JSON, but due to the different capabilities of BioC and PubAnnotation, some issues remain and an improvement on the flexibility of the first conversion tool is desirable. For this reason, we would like to address the following problems in Biomedical Linked Annotation Hackathon 3 (BLAH 3).

- How to revise BioC and PubAnnotation data structures to convey the same information: BioC DTD defines *annotation* and *relation*, and supplementary information is shown using *infon* tags. There is no restriction on *infon*, hence BioC is very flexible. PubAnnotation, on the other hand, uses *denotations* for annotation information. In addition, *relations* always have two entities with subject and object relations, i.e. it does not allow the representation of undirected and n-ary relations. Our goal is to come up with a format that does not lose any information when BioC is converted to PubAnnotation and back again.

- Implementing BioC to PubAnnotation and PubAnnotation to BioC conversion tools: Following a potential agreement between BioC and PubAnnotation organizers, we plan to improve the current BioC-PubAnnotation conversion tool, hence the PubAnnotation system can use BioC files seamlessly or vice versa.

- How to share PubMed and PMC documents in BioC and PubAnnotation: The current BioC repository for PubMed and PMC documents keeps the original XML information as much as possible, however the same set in PubAnnotation only contains minimal text. For example, the BioC PMC set includes

> table/figure captions as well as references, but the PubAnnotation PMC set discards these and carries main text only. This can be a major hurdle to sharing a set of PubMed/PMC documents. Ideally, one should be able to get the same data in both BioC and PubAnnotation PubMed/PMC repositories.

BioC and PubAnnotation have sizeable and growing communities. A unified data structure may be ideal, however having a solution to include the same data while keeping the backward compatibility is a reasonable compromise and may have a bigger impact to the communities.

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

References

1. Comeau, D.C., Islamaj Doğan, R., Ciccarese, P., Cohen, K.B., Krallinger, M., Leitner, F., Lu, Z., Peng, Y., Rinaldi, F., Torii, M., Valencia, A., Verspoor, K., Wiegers, T.C., Wu, C.H., Wilbur, W.J.: BioC: a minimalist approach to interoperability for biomedical text processing. Database **2013**, bat064 (2013)
2. JSON. https://en.wikipedia.org/wiki/JSON
3. Kim, J.-D., Wang, Y.: PubAnnotation: a persistent and sharable corpus and annotation repository. In: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing, pp. 202–205 (2012)
4. PubAnnotation annotation format. http://www.pubannotation.org/docs/annotation-format
5. Comeau, D.C., Doğan, R.I., Kim, S., Wei, C.-H., Wilbur, W.J., Lu, Z.: BioCconvert: a conversion tool between BioC and PubAnnotation. In: International Conference on Biological Ontology & BioCreative 2016 (2016)