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

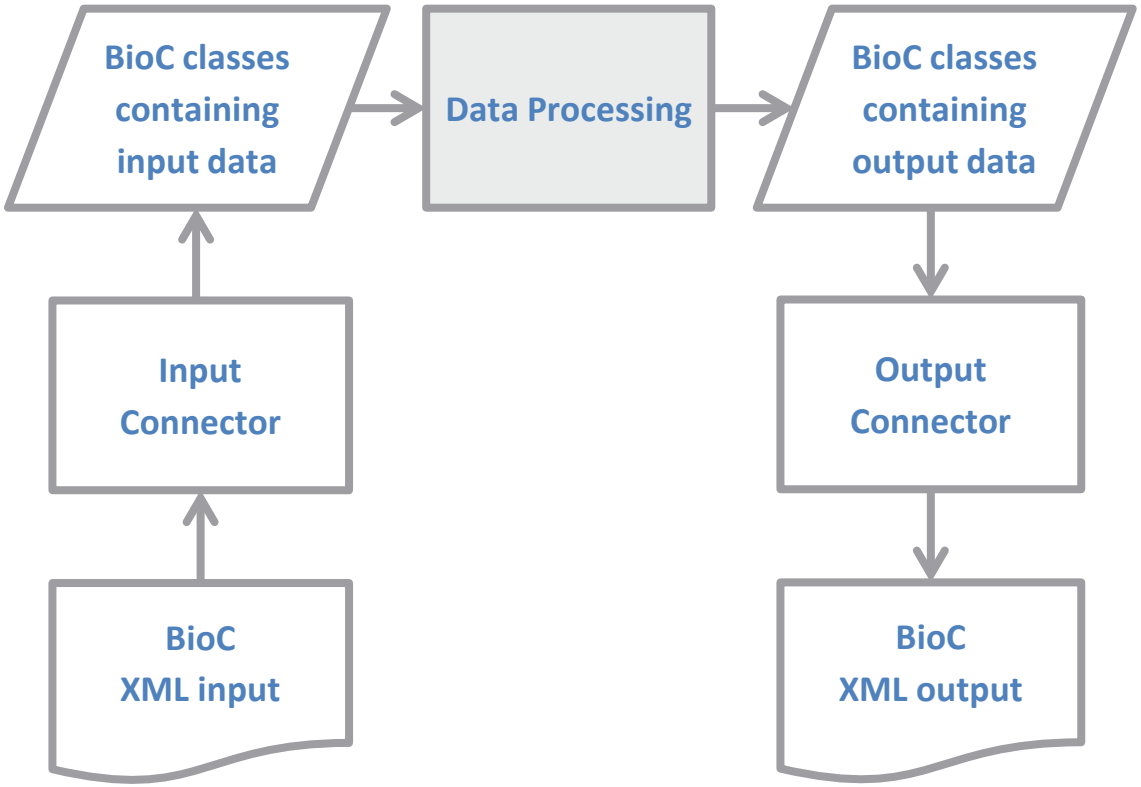
Sun Kim, Donald C. Comeau, Rezarta I. Doğan and Zhiyong Lu

NCBI, NLM, NIH

Biomedical Linked Annotation Hackathon 3 (BLAH3)
Tokyo, Japan
January 16, 2017

# BioC

- A simple approach to interoperability for biomedical text




- XML or JSON
- Available libraries to read/write data
- Little investment to learn new format
- Reduced burden of sharing data and results
- Large repository of datasets in this format

## BioC Implementations

C++
C#
Java
SWIG-Python
SWIG-Perl
Python
Ruby
Go

## BioC Tools

Natural Language Tools:
- Sentence segmenting
- Tokenizing
- Part-of-speech tagging
- Lemmatization
- Dependency parsing

NER Tools:
- Diseases
- Mutations
- Chemicals
- Species
- Genes/proteins

Manual annotation
Sentence simplification
Semantic role labelling
Abbrev. identification

## BioC Corpora

BioC-BioGRID
GeneTag
PMC-BioC
Disease NER
BioNLP Shared task
Abbrev. definition
WBI repository
SRL
iSimp
Metabolites

Comeau et al., Database (Oxford);2013:bat064
Comeau et al., Database (Oxford);2014:bau053

# PubAnnotation

- Repository of text annotation: Share your annotation with others!

- JSON annotation format for communication
- Text alignment function
- Programmable API
- PubMed and PubMed Central

# PubAnnotation Corpora

- 27 released sets and many more beta/developing projects

- PubMedHPO: Human phenotype annotation, based on HPO ontology
- DisGenNET: Disease-Gene association annotation.
- NERUROSES: Cognitive enhancers and anti-depressants drug mentions
- FSU-PRGE: Gene and protein mentions
- PennBioIE: Inhibition of the cytochrome P450 family (CYP450)

# BioC and PubAnnotation

- Are these formats interchangeable?

# 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>
```

# 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>
```

# 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>
```

# 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." }
]
```

# 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." }
]
```

# 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." }
]
```

# 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." }
]
```

# BioC - PubAnnotation

```
<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>
```

```
[
 { "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/source
id/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." }
 ]
```

# BioC - PubAnnotation

```
<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>
```

```
[
 { "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/source
id/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." }
 ]
```

# Other discrepancies

- BioC
  - Multi-span annotation

```
<!ELEMENT annotation ( infon*, location*, text ) >
<!ATTLIST annotation id CDATA #IMPLIED >
<!ELEMENT location EMPTY>
<!ATTLIST location offset CDATA #REQUIRED >
<!ATTLIST location length CDATA #REQUIRED >
```

    - e.g. transmission (TEM) electron microscopy
  - N-ary relations

```
<!ELEMENT relation ( infon*, node* ) >
<!ATTLIST relation id CDATA #IMPLIED >
<!ELEMENT node EMPTY>
<!ATTLIST node refid CDATA #REQUIRED >
<!ATTLIST node role CDATA "" >
```

# Questions

- How to make BioC and PubAnnotation formats 100% compatible?
  - Exception: multi-span annotations and n-ary relations
  - Improve BioC$_{convert}$ (BioC-PubAnnotation conversion tool)

# Questions

- How to share PubMed and PubMed Central documents
  - Version control?
  - Table and figure captions
  - References

# Acknowledgments


- NCBI Text Mining Group
  - Zhiyong Lu (PI)
  - John Wilbur
  - Don Comeau
  - Rezarta Dogan
  - Nicolas Fiorini
  - Alan Hsu
  - Won Kim
  - Robert Leaman
  - Wanli Liu
  - Yifan Peng
  - Natalie Xie
  - Chih-Hsuan Wei
  - Lana Yeganova
- Special Thanks to BLAH3 Organizers!

NIH Intramural Research Program
*Our Research Changes Lives*