BIOC: A MINIMALIST APPROACH TO INTEROPERABILITY FOR BIOMEDICAL TEXT

Rezarta Islamaj Doğan
National Center for Biotechnology Information
National Library of Medicine
• Background on Biocuration and BioCreative
• The problem of interoperability
• The BioC solution for biomedical text processing
• BioC developments
  • BioC libraries
  • Publically available corpora and tools
• Thoughts on the future
The era of accelerated information growth

- Scientists increasingly depend on the availability of each others' data
- Produced at great effort and expense

- Data are only as useful as researchers' ability to locate, integrate and access them
Ref: Hirschman et al. A MOD(ern) perspective on literature curation. Mol Genet Genomics. 2010
The challenge for “biocurators”:

- Extract information from published literature
- Connect information from different sources in a coherent way
- Develop and manage structured vocabularies
- Integrate knowledge bases to represent complex systems
- Make information available
BioCreative

- Critical Assessment of Information Extraction systems in Biology
- Five workshops since 2003
- Shared tasks:
  - Gene mention
  - Gene normalization
  - Protein-protein interaction
  - Document triage
  - Interactive annotation
  - GO annotations
The Problems

- Many different formats
- Many new projects start over
- An atmosphere of competition
The Needs?

• Common format
• Simple-to-learn software to access the format
• Sufficient resources to motivate users

Solution

• a convenient format to share text documents and annotations,
• an input/output library to promote interoperability of data and tools.
THE BIOC APPROACH

Simplicity of use
Little investment to learn a new format
Reduce the burden of sharing

BioC classes containing input data

Input Connector

BioC XML input

Data Processing

BioC classes containing output data

Output Connector

BioC XML output
BioC

- Syntax is XML defined by a DTD
- Software encapsulating data access
- Plain text semantics
BioC DTD

- **Starting point:** a collection of documents
- **Documents:** series of passages
- **Passages:** text or series of sentences
- **Sentences:** text
- **Annotations, Relations**
- **Infon:** basic unit of information
This key file describes the contents of the BioC XML file exampleCollection.xml.

collection: This collection is a simple two-sentence excerpt from an arbitrary PMC article (PMC3048155).

source: PMC (ASCII)

date: yyyyymmdd. Date this example was created.

key: This file

document: This collection contains one document.

id: PubMed Central ID

passage: The first two sentences of the abstract

infon type: paragraph

offset: Article arbitrarily starts at 0.

text: The passage text as it appears in the original document.
Tat mostly activated the MIP-1alpha expression in a p65-dependent manner.

PMID: 22187158
Tat mostly activated the **MIP-1alpha** expression in a **p65**-dependent manner.
Tat mostly activated the MIP-1alpha expression in a p65-dependent manner.

<annotation id="T0">
  <infon key="trigger">Positive_regulation</infon>
  <location offset="11" length="9" />
  <text>activated</text>
</annotation>

<annotation id="T1">
  <infon key="trigger">Gene_expression</infon>
  <location offset="36" length="10" />
  <text>expression</text>
</annotation>

<annotation id="T2">
  <infon key="trigger">Positive_regulation</infon>
  <location offset="55" length="10" />
  <text>dependent</text>
</annotation>
Tat mostly activated the **MIP-1alpha expression** in a p65-dependent manner.

E1: Gene expression / Trigger: expression / Theme: MIP-1alpha

```xml
<relation id="R0">
  <infon key ="event-type">Gene_expression</infon>
  <node refid="G1" role="Theme"/>
  <node refid="T1" role="Trigger"/>
</relation>
```
Tat mostly activated the MIP-1alpha expression in a p65-dependent manner.

**E1:** Gene expression / Trigger: expression / Theme: MIP-1alpha

**E2:** Positive regulation / Trigger: activated / Theme: E1 / Cause: Tat

```xml
<relation id="R1">
  <infon key="event-type">Positive_regulation</infon>
  <node refid="R0" role="Theme"/>
  <node refid="T0" role="Trigger"/>
  <node refid="G0" role="Cause"/>
</relation>
```
Tat mostly activated the MIP-1alpha expression in a p65-dependent manner.

E1: Gene expression / Trigger: expression / Theme: MIP-1alpha

E2: Positive regulation / Trigger: activated / Theme: E1 / Cause: Tat

E3: Positive regulation / Trigger: dependent / Theme: E2 / Cause: p65
Tat mostly activated the MIP-1alpha expression in a p65-dependent manner.
THE BIOC APPROACH

- a convenient format to share text documents and annotations,
- an input/output library to promote interoperability of data and tools.

What can I do with BioC?
<table>
<thead>
<tr>
<th>BioC Implementations</th>
<th>BioC Tools</th>
<th>BioC Corpora</th>
</tr>
</thead>
<tbody>
<tr>
<td>C++</td>
<td>Natural Language Tools: sentence segmenting, tokenizing, part-of-speech tagging, lemmatization, dependency parsing</td>
<td>PMC-BioC</td>
</tr>
<tr>
<td>Java</td>
<td></td>
<td>Disease NER</td>
</tr>
<tr>
<td>SWIG-Python</td>
<td>NER</td>
<td>BioNLP Shared task</td>
</tr>
<tr>
<td>SWIG-Perl</td>
<td>diseases</td>
<td>Abbrev. definition</td>
</tr>
<tr>
<td>Python</td>
<td>mutations</td>
<td>WBI repository</td>
</tr>
<tr>
<td>Ruby</td>
<td>chemicals</td>
<td>SRL</td>
</tr>
<tr>
<td>Go</td>
<td>species</td>
<td>iSimp</td>
</tr>
<tr>
<td></td>
<td>genes/proteins</td>
<td>Metabolites</td>
</tr>
<tr>
<td></td>
<td>Manual annotation</td>
<td></td>
</tr>
</tbody>
</table>
NLP pre-processing pipeline

Incorporates MedPost and Stanford NLP tools and BioLemmatizer

NER tools

- DNorm: disease mention normalization tool to MeSH/OMIM ids
- tmVar: mutation mention identifier
- SR4GN: Species mention and normalization
- tmChem: Chemical names mentions
- GenNorm: Gene Mentions normalized to Entrez Gene

- PubTator: Annotation tool for biomedical concepts

The PMC-BioC Corpus

- PMC Open Access articles in BioC format
- The largest BioC corpus yet

- More than 700,000 full-text articles
- more than 3000 journals

- Why do we need to convert from PMC XML format?

Other Corpora in BioC

**Disease NER:** ~800 PubMed abstracts manually annotated for disease entity mention and concept

**BioNLP Shared Task Corpus:** 6 event extraction tasks cancer genetics, pathway curation and gene regulation

**Abbreviation definition:** more than 3650 PubMed abstracts manually annotated for abbreviations and their definitions in biomedical text

**WBI repository:** a collection of gene, mutation, chemical, protein-protein interactions, disease-treatment, gene expression and other event annotated corpora
Success Stories

- BioCreative IV
  - Interoperability track
  - Gene Ontology (GO) curation task
  - Interactive Curation task (IAT)
  - Comparative Toxicogenomics Database (CTD) Curation task

- BioNLP 2013 shared task contributed resource
CTD success in BioCreative IV

Web service-based NER logical design.

Wiegers T C et al. Database 2014;2014:bau050

© The Author(s) 2014. Published by Oxford University Press.
CTD success in BioCreative IV cont.

BioC-based high-level inter-process communications.

Wiegers T C et al. Database 2014;2014:bau050

© The Author(s) 2014. Published by Oxford University Press.
Paradigm of Cooperation

- Competition motivates
- Cooperation leads to specialization and greater sophistication

BioCreative V BioC task
Thanks: BioC committee

- Paolo Ciccarese, MIND Informatics, Massachusetts General Hospital, Harvard Medical School
- Kevin Bretonnel Cohen, University of Colorado School of Medicine
- Donald C. Comeau, National Center for Biotechnology Information
- Martin Krallinger, Spanish National Cancer Research Centre
- Lynette Hirschman, The MITRE Corporation
- Rezarta Islamaj Doğan, National Center for Biotechnology Information
- Florian Leitner, Spanish National Cancer Research Centre
- Zhiyong Lu, National Center for Biotechnology Information
- Yifan Peng, University of Delaware Center for Bioinformatics & Computational Biology
- Fabio Rinaldi, University of Zurich
- Manabu Torii, University of Delaware Center for Bioinformatics & Computational Biology
- Alfonso Valencia, Spanish National Cancer Research Centre
- Karin Verspoor, National ICT Australia
- Thomas C. Wiegers, Department of Biology at North Carolina State University
- W. John Wilbur, National Center for Biotechnology Information
- Cathy H. Wu, University of Delaware Center for Bioinformatics & Computational Biology
Thanks: BioCreative IV teams

- **PyBioC: a python implementation of the BioC core.** Hernani Marques and Fabio Rinaldi
- **Enhancing the Interoperability of iSimp by Using the BioC Format.** Yifan Peng, Catalina O Tudor, Manabu Torii, Cathy H Wu and K Vijay-Shanker.
- **Improving Interoperability of Text Mining Tools with BioC.** Ritu Khare, Chih-Hsuan Wei, Yuqing Mao, Robert Leaman and Zhiyong
- **Finding Abbreviations in Biomedical Literature: Three BioC-Compatible Modules and Three BioC-formatted Corpora.** Rezarta Islamaj Doğan, Donald C. Comeau, Lana Yeganova and W. John Wilbur.
- **Extending BioC Implementation to More Languages.** Wanli Liu, Donald C. Comeau, Rezarta Islamaj Doğan and W. John Wilbur.
- **Natural Language Processing Pipelines to Annotate BioC Collections with an Application to the NCBI Disease Corpus.** Donald C. Comeau, Haibin Liu, Rezarta Islamaj Doğan and W. John Wilbur.
- **Brat2BioC: conversion tool between brat and BioC.** Antonio Jimeno Yepes, Mariana Neves and Karin Verspoor.
- **A Biomedical Semantic Role Labeling BioC Module for BioCreative IV.** Po-Ting, Lai, Hong-Jie Dai, Johnny Chi-Yang Wu and Richard Tzong-Han Tsai.
- **NaCTeM’s BioC Modules and Resources for BioCreative IV.** Rafal Rak, Riza Batista-Navarro, Andrew Rowley, Makoto Miwa, Jacob Carter and Sophia Ananiadou.
- **Web services-based text mining demonstrates broad impacts for interoperability and process simplification.** Thomas C. Wiegers, Allan Peter Davis and Carolyn J. Mattingly.
- **The Gene Ontology Task at BioCreative IV.** Yuqing Mao, Kimberly Van Auken, Donghui Li, Cecilia N. Arighi and Zhiyong Lu.
- **BioQRator: a web-based interactive biomedical literature curating system.** Dongseop Kwon, Sun Kim, Soo-Yong Shin and W. John Wilbur.
BioC Link