BioC: a minimalist approach to interoperability for biomedical text processing

Don Comeau

Outline

- Background and origin of BioC
- What is BioC?
- Available Tools and Corpora

BioCreative

- Critical Assessment of Information Extraction systems in Biology
- Five workshops since 2004
- Shared tasks:
  - Gene mention
  - Gene normalization
  - Protein-protein interaction
  - Document triage
  - Interactive annotation
  - GO annotations

The problem

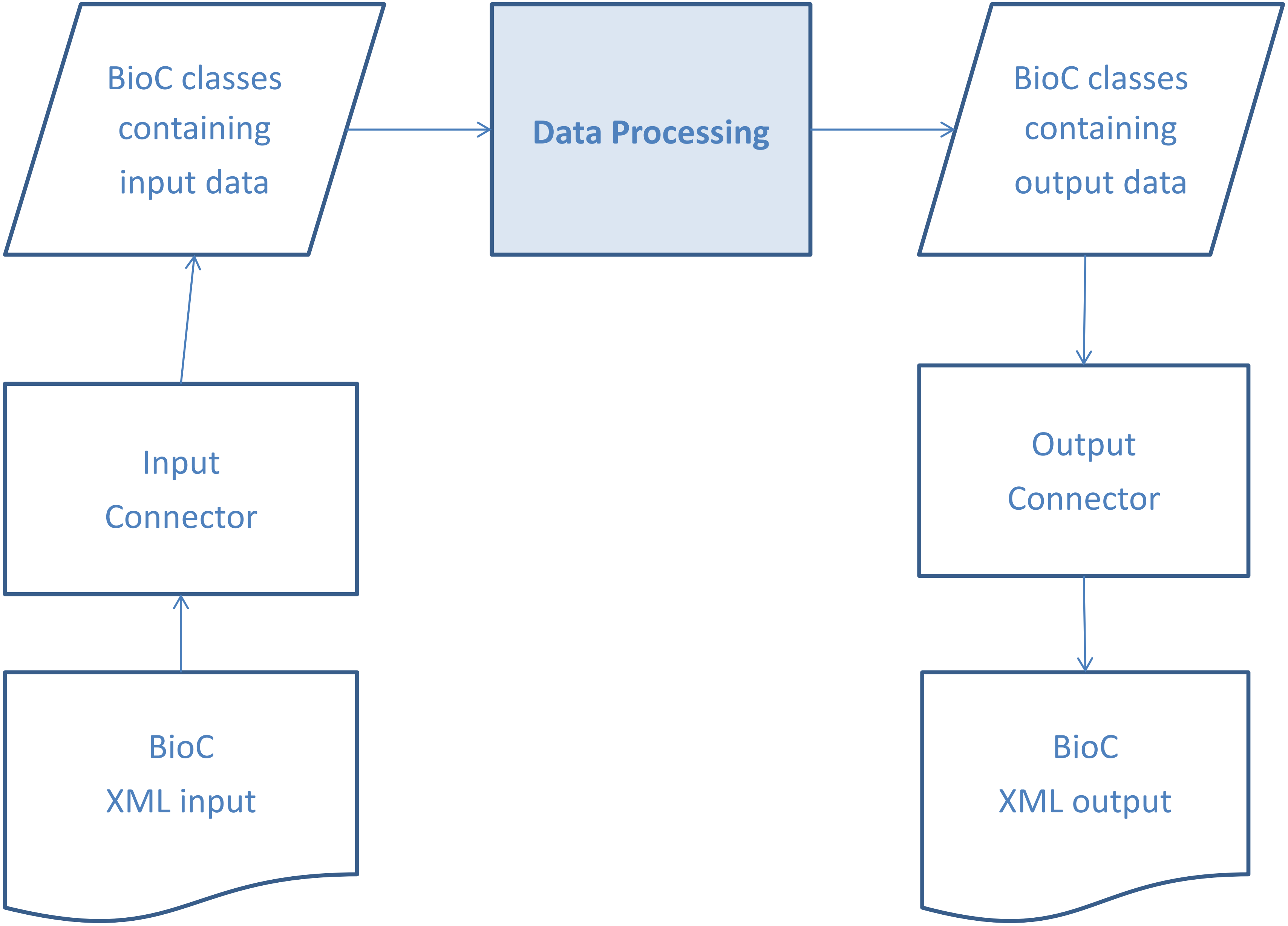
- Many research groups
- Many local data formats
- Many tools
- Hard to build on external tools

Objectives

- Simplicity
- Interoperability
- Broad use and reuse

BioC

- Data format
  - XML DTD
- Code to read and write data
  - Data directly available



BioC classes containing input data

Data Processing

BioC classes containing output data

Input Connector

Output Connector

BioC XML input

BioC XML output

File format

- XML:
  - Easily written and read
  - Portable
  - Familiar

BioC DTD

```
<!ELEMENT collection ( source, date, key, infon*, document+ ) >
<!ELEMENT source (#PCDATA)>
<!ELEMENT date (#PCDATA)>
<!ELEMENT key (#PCDATA)>
<!ELEMENT infon (#PCDATA)>
<!ATTLIST infon key CDATA #REQUIRED >
<!ELEMENT document ( id, infon*, passage+, relation* + ) >
<!ELEMENT id (#PCDATA)>

<!ELEMENT passage( infon*, offset, ((text?, annotation*) | sentence*), relation* ) >
<!ELEMENT offset (#PCDATA)>
<!ELEMENT text (#PCDATA)>

<!ELEMENT sentence ( infon*, offset, text?, annotation*, relation* ) >

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

<!ELEMENT relation ( infon*, node* ) >
```

Starting point:
collection of documents

Documents:
Series of passages

Passage:
text

Passage:
Series of sentences

exampleCollection.xml

```
<!DOCTYPE collection SYSTEM "BioC.dtd">
<collection>
  <source>PubMed Central</source>
  <date>20130123</date>
  <key>exampleCollection.key</key>
  <document>
    <id>PMC3048155</id>
    <passage>
      <infon key="type">paragraph</infon>
      <offset>0</offset>
      <text>The efficacy of computed tomography (CT) screening for early lung cancer detection in heavy smokers is currently being tested by a number of randomized trials. Critical issues remain the frequency of unnecessary treatments and impact on mortality, indicating the need for biomarkers of aggressive disease.</text>
    </passage>
  </document>
</collection>
```

BioC DTD (relations)

```
<!ELEMENT collection ( source, date, key, infon*, document+ ) >
…

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

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

Annotation:
information

Annotation:
text

Relation:
Annotations or relations

exampleAnnotation.xml

```
<!DOCTYPE collection SYSTEM "BioC.dtd">
<collection>
  <source>PubMed Central</source>
  <date>20130123</date>
  <key>exampleAnnotation.key</key>
  <document>
    <id>PMC3048155</id>
    <passage>
      <infon key = "type">paragraph</infon>
      <offset>0</offset>
      <sentence>
        <offset>0</offset>
        <annotation id = "0">
          <infon key = "type">disease name</infon>
          <infon key = "MeSH">D008175</infon>
          <location offset = "61" length = "11" />
          <text>lung cancer</text>
        </annotation>
      </sentence>
    </passage>
  </document>
</collection>
```

exampleAnnotation.xml

```
<!DOCTYPE collection SYSTEM "BioC.dtd">
<collection>
  <source>PubMed Central</source>
  <date>20130123</date>
  <key>exampleAnnotation.key</key>
  <document>
    <id>PMC3048155</id>
    <passage>
```

```
<annotation id = "0">
        <infon key = "type">disease name</infon>
        <infon key = "MeSH">D008175</infon>
        <location offset = "61" length = "11" />
        <text>lung cancer</text>
</annotation>
```

```
      </sentence>
    </passage>
  </document>
</collection>
```

Possible annotations

The efficacy of computed tomography (CT) screening for early lung cancer detection in heavy smokers is currently being tested by a number of randomized trials.

Possible annotations

The efficacy of computed **tomography** (CT) screening for early lung cancer detection in heavy smokers is currently being tested by a number of randomized trials.

id	infon	location		text	Comments
	key:value	offset	length		
T4	PartOfSpeech:NN	25	10	tomography	Part of speech tagging

Possible annotations

The efficacy of computed tomography (CT) screening for early lung cancer detection in heavy **smokers** is currently being tested by a number of randomized trials.

id	infon	location		text	Comments
	key:value	offset	length		
T4	PartOfSpeech:NN	25	10	tomography	Part of speech tagging
L14	lemma:smoker	92	7	smokers	Lemmatization of token

Possible annotations

The efficacy of **computed tomography** (**CT**) screening for early lung cancer detection in heavy smokers is currently being tested by a number of randomized trials.

id	infon	location		text	Comments
	key:value	offset	length		
T4	PartOfSpeech:NN	25	10	tomography	Part of speech tagging
L14	lemma:smoker	92	7	smokers	Lemmatization of token
A1	ABRV:Long Form	16	19	computed tomography	Abbreviation definition in text
A2	ABRV:Short Form	37	2	CT	Abbreviation in text

Possible annotations

The efficacy of computed tomography (CT) screening for early **lung cancer** detection in heavy smokers is currently being tested by a number of randomized trials.

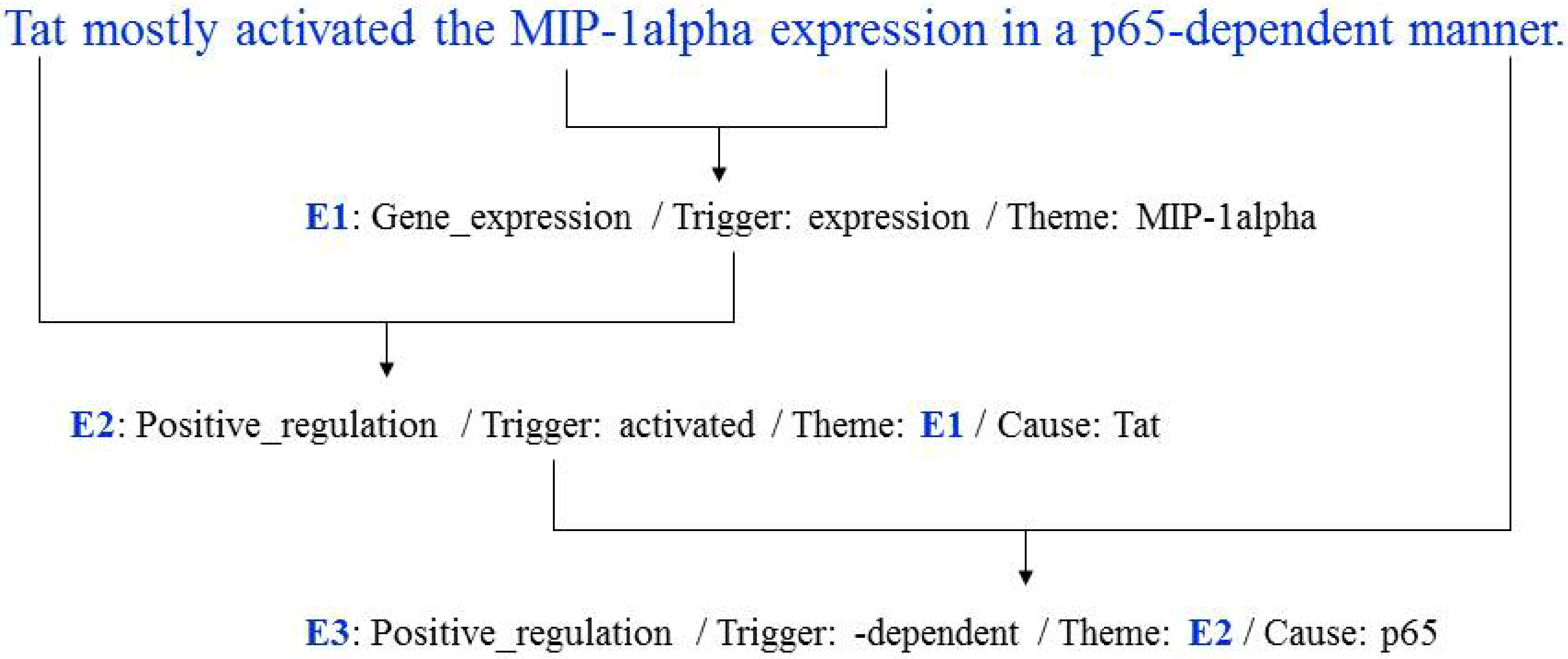
id	infon	location		text	Comments
	key:value	offset	length		
T4	PartOfSpeech:NN	25	10	tomography	Part of speech tagging
L14	lemma:smoker	92	7	smokers	Lemmatization of token
A1	ABRV:Long Form	16	19	computed tomography	Abbreviation definition in text
A2	ABRV:Short Form	37	2	CT	Abbreviation in text
D1	type:disease MeSH:D008175	61	11	lung cancer	Disease name mention in text Concept in terminology resource

Possible annotations

The efficacy of **computed tomography** (CT) **screening** for early lung cancer detection in heavy smokers is currently being tested by a number of randomized trials.

id	infon	location		text	Comments
	key:value	offset	length		
T4	PartOfSpeech:NN	25	10	tomography	Part of speech tagging
L14	lemma:smoker	92	7	smokers	Lemmatization of token
A1	ABRV:Long Form	16	19	computed tomography	Abbreviation definition in text
A2	ABRV:Short Form	37	2	CT	Abbreviation in text
D1	type:disease MeSH:D008175	61	11	lung cancer	Disease name mention in text Concept in terminology resource
E1	type:event	16 41	19 9	computed tomography … screening	Segmented mention annotation

(PMID: 22187158):



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.

```
<annotation id =”G0”>
   <infon key=”type”>Gene_name</infon>
   <location offset=”0” length=”3” />
   <text>Tat</text>
</annotation>

<annotation id =”G1”>
   <infon key=”type”>Gene_name</infon>
   <location offset=”25” length=”10” />
   <text>MIP-1alpha</text>
</annotation>

<annotation id =”G2”>
   <infon key=”type”>Gene_name</infon>
   <location offset=”52” length=”3” />
   <text>p65</text>
</annotation>
```

```
<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.

```
<relation id="R0">
   <infon key =”event-type”>Gene_expression</infon>
   <node refid="G1" role="Theme"/>                  MIP-1alpha
   <node refid="T1" role="Trigger"/>                expression
</relation>

<relation id="R1">
   <infon key =”event-type”>Positive_regulation</infon>
   <node refid="R0" role="Theme"/>
   <node refid="T0" role="Trigger"/>                activated
   <node refid="G0" role="Cause"/>                  Tat
</relation>

<relation id="R2">
   <infon key =”event-type”>Positive_regulation</infon>
   <node refid="R1" role="Theme"/>
   <node refid="T2" role="Trigger"/>                -dependent
   <node refid="G2" role="Cause"/>                  p65
</relation>
```

Semantics

- Not prescribed by BioC
- No way to predict all uses and applications
- Specified in keyfile
- Standard task, use existing keyfile

exampleCollection.key

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: yyyymmdd. 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 from the original document.

Abbreviation key file

annotation: Abbreviations

id: sequential integers from 0 prefixed by either 'SF' or 'LF'

infon["**type**"]: "**ABBR**"

infon["**ABBR**"]: "**ShortForm**" or "**LongForm**"

location: offset: A document offset to where the annotated text begins in the passage or sentence.

length: The length of the annotated text.

text: Original text of the short form or long form.

relation: Long form / short form pair

id: sequential integers from 0 prefixed by 'R‘

infon["**type**"]: "**ABBR**"

node:

role: "**ShortForm**" or "**LongForm**"

refid: id of the appropriate annotation

Implementation

- Clear division between:
  - BioC data classes
  - connector classes to read/write the data (via an XML parser)
  - application code.
- Reading and writing data:
  - Fit entire corpus into memory at once, or
  - Process documents one by one

```
class Node {
  // id of Relation or Annotation
  string refid;
  string role;
};

class Relation {
  string id;
  map<string,string> infons;
  vector<Node> nodes;
};

class Location {
  int offset;
  int length;
};

class Annotation {
  string id;
  map<string,string> infons;
  vector<Location> locations;
  string text;
};
```

```
class Sentence {
  map<string,string> infons;
  int offset;
  string text;
  vector<Annotation> annotations;
};

class Passage {
  map<string,string> infons;
  int offset;
  string text;
  vector<Sentence> sentences;
  vector<Annotation> annotations;
};

class Document {
  string id;
  map<string,string> infons;
  vector<Passage> passages;
};

class Collection {
  string corpus;
  int date;
  string key;
  map<string,string> infons;
  vector<Document> documents;
};
```

BioCreative IV Track 1

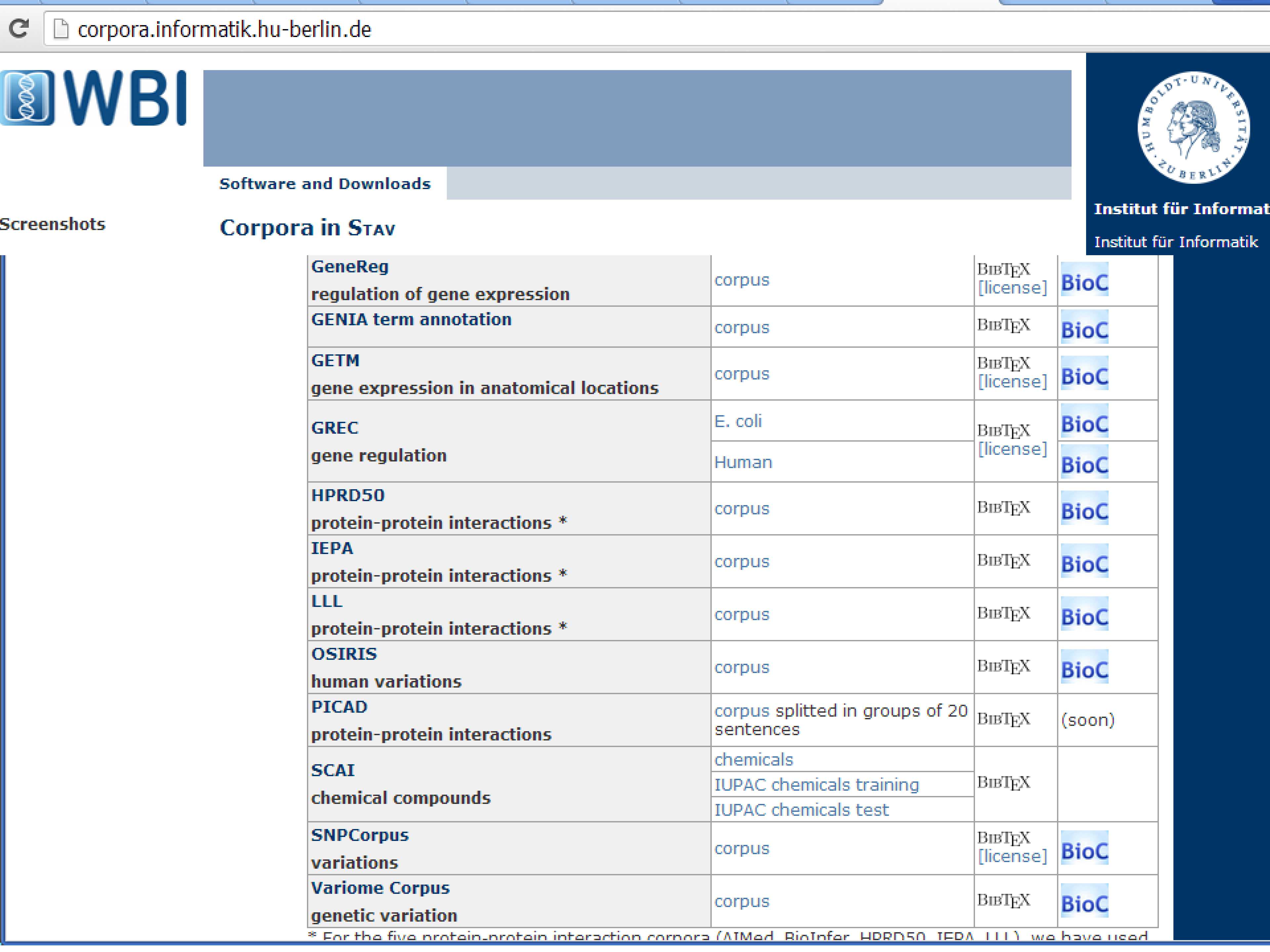
- Interoperability track in BioCreative IV invited participants to contribute new NLP modules to the BioC environment
- 9 accepted papers

Implementations

- C++
- Java (2)
- Python (2)
- Perl
- Go
- Ruby

Corpora

- Abbreviation
  - Ab3P, BIOADI, old Medstract, Schwartz & Hearst
- Disease
- BioNLP Shared Task (4)
- Human Variome Project
- iSimp
- Metabolites
- GO, PMC
- WBI repository (18 corpora)



corpora.informatik.hu-berlin.de

WBI

Software and Downloads

Screenshots

Institut für Informat

Institut für Informatik

Corpora in Stav

GeneReg **regulation of gene expression**	corpus	BibTeX [license]	BioC
GENIA term annotation	corpus	BibTeX	BioC
GETM **gene expression in anatomical locations**	corpus	BibTeX [license]	BioC
GREC **gene regulation**	E. coli	BibTeX [license]	BioC
	Human		BioC
HPRD50 **protein-protein interactions ***	corpus	BibTeX	BioC
IEPA **protein-protein interactions ***	corpus	BibTeX	BioC
LLL **protein-protein interactions ***	corpus	BibTeX	BioC
OSIRIS **human variations**	corpus	BibTeX	BioC
PICAD **protein-protein interactions**	corpus splitted in groups of 20 sentences	BibTeX	(soon)
SCAI **chemical compounds**	chemicals	BibTeX	
	IUPAC chemicals training		
	IUPAC chemicals test		
SNPCorpus **variations**	corpus	BibTeX [license]	BioC
Variome Corpus **genetic variation**	corpus	BibTeX	BioC

* For the five protein-protein interaction corpora (AIMed, BioInfer, HPRD50, IEPA, LLL), we have used

Conversions

- BioNLP Shared Task
- brat
- PubTator
- Argo

Tools

- Sentence segmenting
- Tokenizing
- Part-of-speech tagging
- Lemmatization
- Dependency parsing
- Syntactic parsing

- Sentence simplifying
- Semantic role labeling

- Abbreviation identification
- Named entity recognition
  - Diseases
  - Mutations
  - Species
  - Chemicals
  - Genes / Proteins
- Manual annotation

Available

- http://bioc.sourceforge.net/
- Online
  - Argo
  - BioC-BIOSMILE
  - iSimp
  - Ontogene
- Download
  - NLP pipelines: C++ and Java
  - Abbreviation: S&H, Ab3P, NatLAb
  - tmBioC
  - brat2BioC

Success Stories

- BioCreative IV
  - Gene Ontology (GO) curation task
  - Interactive Curation task (IAT)
  - Comparative Toxicogenomics Database (CTD) Curation task
- BioNLP 2013 shared task contributed resource

CTD Story

- BioCreative III Track CTD Triage
- Impressive results
- Little direct benefit to CTD
- Did not easily integrate into existing pipeline
- BioCreative IV CTD Track
  - Web service
  - BioC format
- Results now useful

Thanks: John Wilbur’s group

- Rezarta Islamaj Doğan
- Sun Kim
- Won Kim
- Haibin Liu
- Wanli Liu
- Natalie Xie
- Lana Yeganova

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

URL

- http://bioc.sourceforge.net/

Addressing the reuse problem

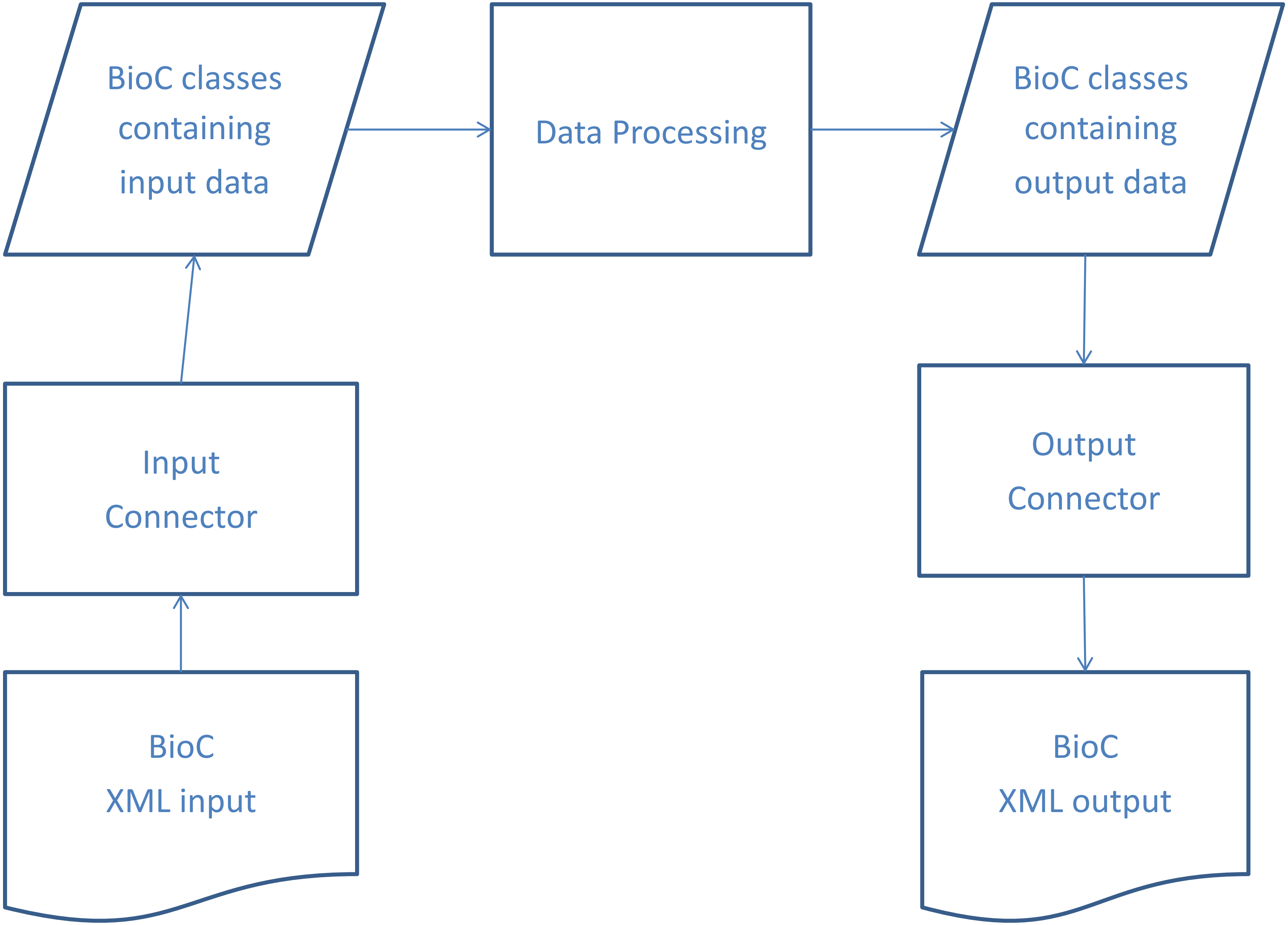
- Object oriented programming
- XML data formatting
- GATE
- UIMA
- GrAF

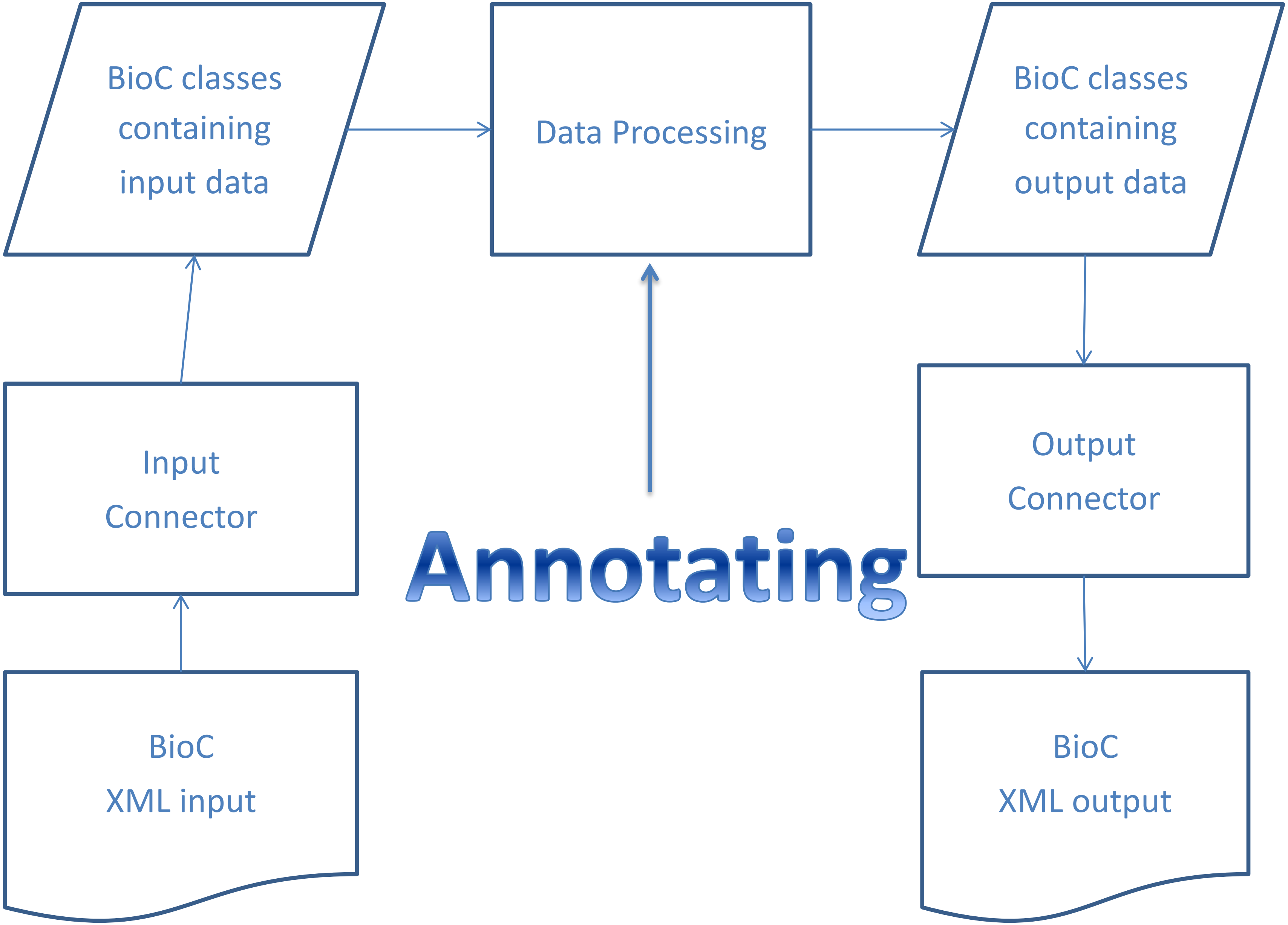
Target audience --- those:

- Developing new techniques
- Using natural language processing
- Producing features for machine learning
- Using text corpora
- Building upon and beyond existing tools

The difference of the new proposal

- Simplicity of use
- There should be little investment to learn to use a format or a software module to process that format
- This will reduce the burden of sharing


BioC classes containing input data
Data Processing
BioC classes containing output data
Input Connector
Output Connector
BioC XML input
BioC XML output



BioC classes containing input data

Data Processing

BioC classes containing output data

Input Connector

Output Connector

BioC XML input

BioC XML output

Annotating

Clinical Data

- BioC can represent clinical text and annotations
  - Based on modest sample of clinical data (2010 i2b2)
  - Based on a few conversations with clinical text researchers

What about other formats?

- BioC is simple
- Does not handle all of the complexity and subtleties of other formats
- Maybe a useful import / export format
- Maybe useful paired with other structured data storage
- Argo (Manchester) works with BioC and UIMA