



<CA>LBC



CALBC Challenge II Guidelines

Table of Contents

Introduction.....	3
Annotations	3
Basic guidelines.....	3
IeXML Schema	4
Inline Annotation.....	4
Further remarks	6
Stand-off Annotation	6
References.....	7
Appendix A – Annotated Text (Inline Annotation Example).....	8
Appendix B - Semantic Groups	12

Introduction

The CALBC project brings together international text mining researchers to address the difficult issue of annotating large text corpora with a large set of semantic types. The task is the annotation of named entities in a large biomedical corpus, for a variety of semantic categories. Participants can download the corpus, can annotate it with their own text mining solutions, submit the corpus to a central server and receive an assessment of their results through a fully automated analysis. The annotated, harmonized corpus (so called Silver Standard) becomes a resource for the community to be used as a reference for improving text mining applications.

A secondary goal of this project is to define a standardized format for representing the annotations contributed by the participants and comparing them effectively. The final corpus will also be made available formatted in RDF for exploitation in Semantic Web applications.

For further information about CALBC in general, see www.calbc.eu and [1], [2], [3].

Annotations

To annotate entity mentions in text, in CALBC the IeXML annotation schema is used [4]. IeXML has been designed as an annotation schema that is easy to understand and work with, and, at the same time, is complete enough to normalize different types of annotations. In Appendix A, an example text is given that has been annotated using the IeXML schema.

Basic guidelines

- Annotations are specified in XML, since it is both machine and human readable.
- Inline annotation is preferred. However, stand-off annotation is also accepted. Examples for both annotation styles are presented below.
- For each entity annotation, a concept identifier together with a namespace should be specified that identifies the annotated concept in an (established) knowledge source.
- For each annotation, a semantic type from the UMLS Semantic Network or a semantic group, as specified in Appendix B, must be provided.
- For each annotation, the exact boundaries should be specified.
- If an annotation system detects nested annotations that belong to the same semantic group, only the annotation with the largest span in text should be accepted. For example, in the case of *lung cancer*, annotate *lung cancer*, but do not additionally annotate *cancer*. Still, the annotation of overlapping entities is allowed provided they belong to different semantic groups. For example, if a system targets both anatomical entities and diseases, it should annotate both, *lung* and *lung cancer*, since they belong to different semantic groups (anatomy, and diseases, respectively).
- Given an annotation, if the annotation system cannot decide between several candidate identifiers of the annotation within a knowledge source, it should provide all of them.

leXML Schema

In this section the leXML annotation schema is described. Appendix A shows an example of annotated text using the leXML schema.

Inline Annotation

Entity annotation

The annotation of entities is done using the *e* element. The *e* element should enclose the text where an entity / entities can be found. While different annotation granularities are conceivable (exact boundaries, noun phrase level, sentence level, etc.), the specification of exact boundaries of annotations is preferred.

Concept identifier specification

Use the *id* attribute of the *e* element to specify the concept identifier of the annotated entity within an established knowledge source (mandatory). It should consist of the namespace of the knowledge source (e.g., “UMLS”) the actual identifier (e.g., “C0001403”), the semantic type of the annotated entity and the corresponding semantic group, separated by colons. The namespace is required, since different knowledge sources are allowed to be used to identify annotated entities within one challenge submission. The semantic type specification will allow us to analyze agreement on the semantic type level, even if different identifiers have been chosen by different participants of the challenge. This is relevant for building the CALBC Silver Standard III. The semantic group assignment is optional, if a semantic type has been provided (and mandatory otherwise).

Examples

This is an example *e* element with *id* attribute, where no semantic group is given:

```
<e id="UMLS:C0001403:T047:">Addison's disease</e>
```

In the case of multiple concept identifiers for the same annotation, the pipe character „|“ is used to separate the identifiers (N1:ID1:ST1:SG1|N2:ID2:ST2:SG2).

In the following example, a piece of text is given that can be annotated with a UniProt entity and a UMLS entity. Four different variants of valid annotation are shown (while variant 1. is the preferred one):

1. `<e id="Uniprot:P01308:T028:PRGE|UMLS:C1337112:T028:PRGE">INS gene</e>`
2. `<e id="Uniprot:P01308:T028:|UMLS:C1337112:T028:">INS gene</e>`
3. `<e id="Uniprot:P01308::PRGE|UMLS:C1337112::PRGE">INS gene</e>`
4. `<e id=":::PRGE">INS gene</e>`

In 1. till 3. both possible concept identifiers are specified. While in 1. all possible pieces of information are specified (namespace, ID, semantic type and semantic group), in 2. no semantic group and in 3. no semantic type is specified. In 4., only a semantic group is

specified. Submission of this type will be considered for the named entity recognition part of the CALBC Challenge II, but not for the concept identification part.

In the remainder of this document in any examples we will omit the semantic group specification.

Token annotation

When entity mentions are not constituted by a continuous stream of text, or when there are overlapping entity mentions, the different pieces of text belonging to the entity mention must be specified. This requires the text to be divided into smaller units (tokens).

There are no particular constraints on how the tokenization is performed, apart from two requirements: Use the *w* element to specify tokens, and provide each token with a unique token identifier. Use the *id* attribute of the *w* element to represent the token identifier (mandatory).

To identify the tokens belonging to an entity annotation, append a colon followed by the comma separated list of the respective token identifiers at the end of the concept identifier in the corresponding *e* element.

The following example deals with a more complicated case:

```
<e id="UMLS:C0222601:T023::1,2|UMLS:C0006142:T191::2,3">
  <w id="1">left</w>
  <w id="2">breast</w>
  <w id="3">cancer</w>
</e>
```

Here, “left breast” (i.e., tokens 1 and 2) is identified by UMLS:C0222601:T023, while “breast cancer” (tokens 2 and 3) is identified by “UMLS:C0006142:T191”, within one *e* element. (Note that there is no concept for “left breast cancer” in the UMLS).

Identifying tokens allows post-coordinated concepts to be represented, even in the absence of overlap. For example, the entity “acute bacterial meningitis” can be annotated with two UMLS concepts: “UMLS:C0205178:T079” for “acute” (token 1) and “UMLS:C0085437:T047” for “bacterial meningitis” (token 2):

```
<e id="UMLS:C0205178:T079::1|UMLS:C0085437:T047::2">
  <w id="1">acute</w>
  <w id="2">bacterial meningitis</w>
</e>
```

The specification of tokens is optional. The two examples above can be represented more simply as follows:

```
<e id="UMLS:C0222601:T023|UMLS:C0006142:T191">left breast cancer</e>

<e id="UMLS:C0205178:T079|UMLS:C0085437:T047">acute bacterial meningitis</e>
```

In the first example concepts with overlapping textual mentions are annotated, other than in the second one.

Further remarks

- Syntactical conformance for the submission of annotations can be checked by using the CALBC DTD (<http://dtd.calbc.eu/>) or Appendix C). Every participant should check the submission using an XML validator referring to this DTD.
- The *id* attribute of both *e* and *w* elements is mandatory. Do not introduce further attributes.
- For each submission, a list of the namespaces used in concept identifiers must be provided, together with the name of the knowledge source they refer to and the version of the knowledge source used.

Stand-off Annotation

In some cases, the proposed inline annotation is not convenient (e.g. when the document cannot be disclosed). Thus a stand-off annotation is preferred. To cope with stand-off annotations, use the following modifications of the inline annotation format:

- Use the *text* element to indicate text areas in the XML document where entities might be found.
- Provide an *id* attribute for each text element.
- For each text area, collect the entity annotations and represent them using the *e* element and its *id* attribute, as described for the inline annotation case.
- Use the *annotation* element to enclose the annotations belonging to a particular text area.
- Specify the text area, to which the annotations belong to using the *text_id* attribute.

In contrast to the inline annotation case, in the stand-off case for each annotation (i.e., in each *e* element) its offset in text and its length must be specified, using the attributes *offset* and *length*. The offset counting starts at the beginning of the specified text area. Both offset and length count refer to characters, not bytes. This is an example for the stand-off annotation format:

```

<document>
  <text id="1">
    Clarifying CB2 receptor-dependent and independent
    effects of THC on human lung epithelial cells.
  </text>
  <annotation>
    <e text_id="1" offset="11" length="3" id="uniprot:P34972:T116">CB2</e>
    <e text_id="1" offset="68" length="5" id="species:3708:T016">human</e>
  </annotation>
</document>

```

The stand-off format still allows the annotation of tokens. In this case, the token annotations are specified as content of the *e* element. The way how tokens are referenced in the *id* attribute of the *e* element remains the same as in the inline annotation case. In the following, the 'left breast cancer' example introduced above is represented in stand-off format:

```

<document>
  <text id="1">
    A 70-year-old woman underwent modified radical mastectomy on her
    left breast cancer and received oral 5-fluorouracil derivatives for 2 years
    in another hospital.
  </text>
  <annotation>
    <e text_id="1" offset="65" length="18"
id="UMLS:C0222601:T023::1,2|UMLS:C0006142:T191::2,3">
      <w id="1">left</w>
      <w id="2">breast</w>
      <w id="3">cancer</w>
    </e>
  </annotation>
</document>

```

References

- [1] Rebholz-Schuhmann,D., A.J. Jimeno Yepes, E.M. van Mulligen, N. Kang, J. Kors, D. Milward, P. Corbett, E. Buyko, K. Tomanek, E. Beisswanger, and U. Hahn. (2010) The CALBC Silver Standard Corpus for Biomedical Named Entities: A Study in Harmonizing the Contributions from Four Independent Named Entity Taggers. *Proc. of the LREC 2010*.
- [2] Rebholz-Schuhmann,D., A.J. Jimeno Yepes, E.M. Van Mulligen, N. Kang, J. Kors, D. Milward, P. Corbett, E. Buyko, E. Beisswanger, and U. Hahn. (2010) "CALBC Silver Standard Corpus." *J Bioinform Comput Biol.* 2010 Feb;8(1):163-79.
- [3] Rebholz-Schuhmann,D., et al. (2010) " Assessment of NER solutions against the first and second CALBC Silver Standard Corpus." *Proc. of the International Symposium on Semantic Mining in Biomedicine 2010* (to appear).
- [4] Rebholz-Schuhmann,D., Kirsch,H., and Nenadic,G. (2006) IeXML: towards a framework for interoperability of text processing modules to improve annotation of semantic types in biomedical text. *Proc. of the BioLINK workshop at ISMB 2006*.

Appendix A – Annotated Text (Inline Annotation Example)

Original Document

```

<PubmedArticle>
<MedlineCitation Owner="NLM" Status="MEDLINE">
<PMID>1410221</PMID>
<DateCreated>
  <Year>1992</Year>
  <Month>11</Month>
  <Day>12</Day>
</DateCreated>
<DateCompleted>
  <Year>1992</Year>
  <Month>11</Month>
  <Day>12</Day>
</DateCompleted>
<DateRevised>
<Year>2004</Year>
<Month>11</Month>
<Day>17</Day>
</DateRevised>
<Article PubModel="Print"><Journal><ISSN IssnType="Print">0033-3506</ISSN><JournalIssue
CitedMedium="Print"><Volume>106</Volume><Issue>5</Issue><PubDate><Year>1992</Year><Month>Sep</Month>
</PubDate></JournalIssue><Title>Public health</Title></Journal><ArticleTitle>Internal variation in
the uptake of whooping cough immunisation within a Health
Authority.</ArticleTitle><Pagination><MedlinePgn>367-
74</MedlinePgn></Pagination><Abstract><AbstractText>Data were obtained on the vaccination history of
6,898 children immunised with D3, the final dose of diphtheria vaccine. These children were born
between 1984 and 1990 and were also resident within a single Health Authority during part or all of
this time. Analysis of this data revealed consistent variation by treatment centre in the uptake of
whooping cough vaccine in those children receiving the diphtheria vaccine. Certain treatment centres
consistently achieved greater success in giving the whooping cough vaccine to those children who had
received the diphtheria vaccine.</AbstractText></Abstract><Affiliation>Department of Public Health
Medicine, North East Warwickshire Health Authority, Nuneaton, Warwickshire.</Affiliation><AuthorList
CompleteYN="Y"><Author
ValidYN="Y"><LastName>Janes</LastName><ForeName>H</ForeName><Initials>H</Initials></Author></AuthorL
ist><Language>eng</Language><PublicationTypeList><PublicationType>Journal
Article</PublicationType></PublicationTypeList></Article><MedlineJournalInfo><Country>ENGLAND</Count
ry><MedlineTA>Public
Health</MedlineTA><NlmUniqueID>0376507</NlmUniqueID></MedlineJournalInfo><ChemicalList><Chemical><Re
gistryNumber>0</RegistryNumber><NameOfSubstance>Diphtheria
Toxoid</NameOfSubstance></Chemical><Chemical><RegistryNumber>0</RegistryNumber><NameOfSubstance>Pert
ussis
Vaccine</NameOfSubstance></Chemical></ChemicalList><CitationSubset>IM</CitationSubset><MeshHeadingLi
st><MeshHeading><DescriptorName MajorTopicYN="N">Child,
Preschool</DescriptorName></MeshHeading><MeshHeading><DescriptorName MajorTopicYN="N">Databases,
Factual</DescriptorName></MeshHeading><MeshHeading><DescriptorName MajorTopicYN="N">Diphtheria
Toxoid</DescriptorName><QualifierName MajorTopicYN="Y">administration &
dosage</QualifierName></MeshHeading><MeshHeading><DescriptorName
MajorTopicYN="N">Humans</DescriptorName></MeshHeading><MeshHeading><DescriptorName
MajorTopicYN="N">Immunization Schedule</DescriptorName></MeshHeading><MeshHeading><DescriptorName
MajorTopicYN="N">Infant</DescriptorName></MeshHeading><MeshHeading><DescriptorName
MajorTopicYN="N">Pertussis Vaccine</DescriptorName><QualifierName MajorTopicYN="Y">administration
& dosage</QualifierName></MeshHeading><MeshHeading><DescriptorName MajorTopicYN="N">Whooping
Cough</DescriptorName><QualifierName
MajorTopicYN="Y">immunology</QualifierName></MeshHeading></MeshHeadingList></MedlineCitation><Pubmed

```

19.10.2010

```
Data<History><PubMedPubDate
PubStatus="pubmed"><Year>1992</Year><Month>9</Month><Day>1</Day></PubMedPubDate><PubMedPubDate
PubStatus="medline"><Year>1992</Year><Month>9</Month><Day>1</Day><Hour>0</Hour><Minute>1</Minute></P
ubMedPubDate></History><PublicationStatus>ppublish</PublicationStatus><ArticleIdList><ArticleId
IdType="pubmed">1410221</ArticleId><ArticleId
IdType="medline">93029182</ArticleId></ArticleIdList></PubmedData></PubmedArticle> -
```

Annotated Document

<?xml version="1.0"?>

<!DOCTYPE PubmedArticleSet PUBLIC "-//CALBC//DTD 07-10-2010//EN"

"http://dtd.calbc.eu/calbc_071010.dtd">

<PubmedArticle><MedlineCitation Owner="NLM"
 Status="MEDLINE"><PMID>1410221</PMID><DateCreated><Year>1992</Year><Month>11</Month><Day>12</Day></D
 ateCreated><DateCompleted><Year>1992</Year><Month>11</Month><Day>12</Day></DateCompleted><DateRevis
 e><Year>2004</Year><Month>11</Month><Day>17</Day></DateRevised><Article
 PubModel="Print"><Journal><ISSN IssnType="Print">0033-3506</ISSN><JournalIssue
 CitedMedium="Print"><Volume>106</Volume><Issue>5</Issue><PubDate><Year>1992</Year><Month>Sep</Month>
 </PubDate></JournalIssue><Title>Public health</Title></Journal><ArticleTitle>Internal variation in
 the uptake of < e id="umls:C0043168:T047">whooping cough</e> immunisation within a Health
 Authority.</ArticleTitle><Pagination><MedlinePgn>367-
 74</MedlinePgn></Pagination><Abstract><AbstractText>< e id="umls:C1511726:T078">Data</e> were < e
 id="umls:C1301820:T169">obtained</e> on the < e
 id="umls:C0042196:T061::1|umls:C0019665:T170::2|umls:C0008059:T100::4"> <w id="1">vaccination</w> <w
 id="2">history</w> <w id="3">of 6,898</w> <w id="4">children</w></e> immunised with D3, the < e
 id="umls:C0205088:T079::1|umls:C0178602:T081::2|umls:C0012551:T109::4<w id="1">final</w> <w
 id="2">dose</w><w id="3">of</w> <w id="4">diphtheria vaccine</w></e>. < e
 id="umls:C0008059:T100">These children</e> were born between 1984 and < e
 id="umls:C0681758:T079">1990</e> and were < e id="umls:C1549439:T170">also resident within a </e> < e
 id="umls:C1549113:T033:1|umls:C1273803:T093::2"> <w id="1">single</w> <w id="2">Health
 Authority</w></e>< e id="umls:C0449719:T082">during part</e> or < e id="umls:C0040223:T079">all of
 this time</e>. < e id="umls:C0010992:T057">Analysis of this data</e> < e
 id="umls:C0443289:T080">revealed</e> < e id="umls:C0332290:T078|umls:C0205419:T080">consistent
 variation</e> < e id="umls:C0039798:T169|umls:C0205099:T082">by treatment centre</e> < e
 id="umls:C0347980:T081|umls:C0031237:T109">in the uptake of whooping cough vaccine</e> < e
 id="umls:C0008059:T100">in those children</e> < e id="umls:C1514756:T080">receiving</e> < e
 id="umls:C0012551:T109">the diphtheria vaccine</e>. < e
 id="umls:C0439543:T080|umls:C0039798:T169|umls:C0205099:T082">Certain treatment centres
 consistently</e> achieved < e id="umls:C0443228:T081|umls:C0597535:T054">greater success</e> in
 giving < e id="umls:C0031237:T109">the whooping cough vaccine</e> < e id="umls:C0008059:T100">to those
 children</e> who had < e id="umls:C1514756:T080">received</e> < e id="umls:C0012551:T109">the
 diphtheria vaccine</e>.</AbstractText></Abstract><Affiliation>Department of Public Health Medicine,
 North East Warwickshire Health Authority, Nuneaton, Warwickshire.</Affiliation><AuthorList
 CompleteYN="Y"><Author
 ValidYN="Y"><LastName>Janes</LastName><ForeName>H</ForeName><Initials>H</Initials></Author></AuthorL
 ist><Language>eng</Language><PublicationTypeList><PublicationType>Journal
 Article</PublicationType></PublicationTypeList></Article><MedlineJournalInfo><Country>ENGLAND</Count
 ry><MedlineTA>Public
 Health</MedlineTA><NlmUniqueID>0376507</NlmUniqueID></MedlineJournalInfo><ChemicalList><Chemical><Re
 gistryNumber>0</RegistryNumber><NameOfSubstance>Diphtheria
 Toxoid</NameOfSubstance></Chemical><Chemical><RegistryNumber>0</RegistryNumber><NameOfSubstance>Pert
 ussis
 Vaccine</NameOfSubstance></Chemical></ChemicalList><CitationSubset>IM</CitationSubset><MeshHeadingLi
 st><MeshHeading><DescriptorName MajorTopicYN="N">Child,
 Preschool</DescriptorName></MeshHeading><MeshHeading><DescriptorName MajorTopicYN="N">Databases,
 Factual</DescriptorName></MeshHeading><MeshHeading><DescriptorName MajorTopicYN="N">Diphtheria
 Toxoid</DescriptorName><QualifierName MajorTopicYN="Y">administration & amp;
 dosage</QualifierName></MeshHeading><MeshHeading><DescriptorName
 MajorTopicYN="N">Humans</DescriptorName></MeshHeading><MeshHeading><DescriptorName
 MajorTopicYN="N">Immunization Schedule</DescriptorName></MeshHeading><MeshHeading><DescriptorName
 MajorTopicYN="N">Infant</DescriptorName></MeshHeading><MeshHeading><DescriptorName
 MajorTopicYN="N">Pertussis Vaccine</DescriptorName><QualifierName MajorTopicYN="Y">administration
 & amp; dosage</QualifierName></MeshHeading><MeshHeading><DescriptorName MajorTopicYN="N">Whooping
 Cough</DescriptorName><QualifierName
 MajorTopicYN="Y">immunology</QualifierName></MeshHeading></MeshHeadingList></MedlineCitation><Pubmed

19.10.2010

```
Data><History><PubMedPubDate
PubStatus="pubmed"><Year>1992</Year><Month>9</Month><Day>1</Day></PubMedPubDate><PubMedPubDate
PubStatus="medline"><Year>1992</Year><Month>9</Month><Day>1</Day><Hour>0</Hour><Minute>1</Minute></P
ubMedPubDate></History><PublicationStatus>ppublish</PublicationStatus><ArticleIdList><ArticleId
IdType="pubmed">1410221</ArticleId><ArticleId
IdType="medline">93029182</ArticleId></ArticleIdList></PubmedData></PubmedArticle> -
```

Appendix B - Semantic Groups

The following table maps the UMLS semantic types to a higher level classification (semantic groups). It is a revised version of the classification available from the NLM¹.

The semantic groups used in CALBC are based on those defined in the UMLS. The following groups are distinguished:

	Abbreviation
Activities and behaviors	ACTI
Anatomy	ANAT
Chemicals and drugs	CHED
Concepts and ideas	CONC
Devices	DEVI
Disorders	DISO
Geographic areas	GEOG
Living beings	LIVB
Objects	OBJC
Occupations	OCCU
Organizations	ORGA
Phenomena	PHEN
Physiology	PHYS
Proteins and genes	PRGE
Procedures	PROC
Miscellaneous	MISC

Table 11: List of the UMLS semantic groups.

The following table maps each semantic type as defined in the UMLS to one of the semantic groups above.

Type	Description	Group
T052	Activity	ACTI
T053	Behavior	ACTI
T056	Daily or Recreational Activity	ACTI
T051	Event	ACTI
T064	Governmental or Regulatory Activity	ACTI
T055	Individual Behavior	ACTI
T066	Machine Activity	ACTI
T057	Occupational Activity	ACTI
T054	Social Behavior	ACTI
T017	Anatomical Structure	ANAT
T029	Body Location or Region	ANAT
T023	Body Part Organ or Organ Component	ANAT
T030	Body Space or Junction	ANAT
T031	Body Substance	ANAT
T022	Body System	ANAT
T025	Cell	ANAT
T026	Cell Component	ANAT
T018	Embryonic Structure	ANAT
T021	Fully Formed Anatomical Structure	ANAT

1 <http://semanticnetwork.nlm.nih.gov/SemGroups/>

T024	Tissue	ANAT
T195	Antibiotic	CHED
T123	Biologically Active Substance	CHED
T122	Biomedical or Dental Material	CHED
T118	Carbohydrate	CHED
T103	Chemical	CHED
T120	Chemical Viewed Functionally	CHED
T104	Chemical Viewed Structurally	CHED
T200	Clinical Drug	CHED
T111	Eicosanoid	CHED
T196	Element Ion or Isotope	CHED
T131	Hazardous or Poisonous Substance	CHED
T125	Hormone	CHED
T129	Immunologic Factor	CHED
T130	Indicator Reagent or Diagnostic Aid	CHED
T197	Inorganic Chemical	CHED
T119	Lipid	CHED
	Neuroreactive Substance or Biogenic	
T124	Amine	CHED
T114	Nucleic Acid Nucleoside or Nucleotide	CHED
T109	Organic Chemical	CHED
T115	Organophosphorus Compound	CHED
T121	Pharmacologic Substance	CHED
T110	Steroid	CHED
T127	Vitamin	CHED
T185	Classification	CONC
T077	Conceptual Entity	CONC
T169	Functional Concept	CONC
T102	Group Attribute	CONC
T078	Idea or Concept	CONC
T170	Intellectual Product	CONC
T171	Language	CONC
T080	Qualitative Concept	CONC
T081	Quantitative Concept	CONC
T089	Regulation or Law	CONC
T082	Spatial Concept	CONC
T079	Temporal Concept	CONC
T203	Drug Delivery Device	DEVI
T074	Medical Device	DEVI
T075	Research Device	DEVI
T020	Acquired Abnormality	DISO
T190	Anatomical Abnormality	DISO
T019	Congenital Abnormality	DISO
T047	Disease or Syndrome	DISO
T050	Experimental Model of Disease	DISO
T048	Mental or Behavioral Dysfunction	DISO
T191	Neoplastic Process	DISO
T083	Geographic Area	GEOG
T100	Age Group	LIVB
T003	Alga	LIVB
T011	Amphibian	LIVB
T008	Animal	LIVB

T194	Archaeon	LIVB
T007	Bacterium	LIVB
T012	Bird	LIVB
T099	Family Group	LIVB
T013	Fish	LIVB
T004	Fungus	LIVB
T096	Group	LIVB
T016	Human	LIVB
T009	Invertebrate	LIVB
T015	Mammal	LIVB
T001	Organism	LIVB
T101	Patient or Disabled Group	LIVB
T002	Plant	LIVB
T098	Population Group	LIVB
T097	Professional or Occupational Group	LIVB
T014	Reptile	LIVB
T006	Rickettsia or Chlamydia	LIVB
T010	Vertebrate	LIVB
T005	Virus	LIVB
T071	Entity	OBJC
T168	Food	OBJC
T073	Manufactured Object	OBJC
T072	Physical Object	OBJC
T167	Substance	OBJC
T091	Biomedical Occupation or Discipline	OCCU
T090	Occupation or Discipline	OCCU
T093	Health Care Related Organization	ORGA
T092	Organization	ORGA
T094	Professional Society	ORGA
T095	Self-help or Relief Organization	ORGA
T038	Biologic Function	PHEN
T069	Environmental Effect of Humans	PHEN
T068	Human-caused Phenomenon or Process	PHEN
T034	Laboratory or Test Result	PHEN
T070	Natural Phenomenon or Process	PHEN
T067	Phenomenon or Process	PHEN
T043	Cell Function	PHYS
T201	Clinical Attribute	PHYS
T045	Genetic Function	PHYS
T041	Mental Process	PHYS
T044	Molecular Function	PHYS
T042	Organ or Tissue Function	PHYS
T032	Organism Attribute	PHYS
T040	Organism Function	PHYS
T039	Physiologic Function	PHYS
T087	Amino Acid Sequence	PRGE
T088	Carbohydrate Sequence	PRGE
T028	Gene or Genome	PRGE
T085	Molecular Sequence	PRGE
T086	Nucleotide Sequence	PRGE
T116	Amino Acid Peptide or Protein	PRGE

T126	Enzyme	PRGE
T192	Receptor	PRGE
T060	Diagnostic Procedure	PROC
T065	Educational Activity	PROC
T058	Health Care Activity	PROC
T059	Laboratory Procedure	PROC
T063	Molecular Biology Research Technique	PROC
T062	Research Activity	PROC
T061	Therapeutic or Preventive Procedure	PROC
T049	Cell or Molecular Dysfunction	MISC
T033	Finding	MISC
T037	Injury or Poisoning	MISC
T046	Pathologic Function	MISC
T184	Sign or Symptom	MISC

Table 12: List of the UMLS Semantic Types used for the project and their mapping to the the UMLS Semantic Groups.

Appendix C – CALBC DTD

<!--

This is the Current DTD which NLM has written for External Use. If you are a NCBI User, use the information from the PubmedArticleSet.

Comments and suggestions are welcome.
(May 9, 2000)

Corrections:

~~~~~

Oct. 09 2002

- "PubMedArticle" has been renamed to "PubmedArticle"
- All referencies to "PubMedArticle" has been removed
- "ProviderId" has been removed from PubmedData
- "URL" has been removed from PubmdeData

\$Id: pubmed\_060101.dtd 70528 2005-10-07 18:51:25Z korobtch \$

-->

<!--

=====

>

<!--

=====

>

<!-- Reference to Where the MEDLINECITATION DTD is located -->  
<!ENTITY % Medline PUBLIC "-//NLM//DTD Medline, 01 Nov 2004//EN"  
"nlmmedline\_060101.dtd">

% Medline;

<!--

=====

>

<!ENTITY % ArticleTitle.Ref "ArticleTitle">  
<!ENTITY % ISSN.Ref "ISSN?">  
<!ENTITY % Pub.Date.Ref "PubDate?">  
<!ENTITY % iso.language.codes "(AF|AR|AZ|BG|CS|DA|DE|EN|EL|ES|FA|FI|FR|HE|  
HU|HY|IN|IS|IT|IW|JA|KA|KO|LT|MK|ML|NL|NO|  
PL|PT|PS|RO|RU|SL|SK|SQ|SR|SV|SW|TH|TR|UK|  
VI|ZH)">

<!ENTITY % pub.status.int "pmc | pmcr | pubmed | pubmedr |  
premedline | medline | medliner">

<!ENTITY % pub.status "(received | accepted | epublish |  
ppublish | revised | aheadofprint |  
retracted | %pub.status.int;)">

<!ENTITY % art.id.type.int "pubmed | medline | pmcid">

<!ENTITY % art.id.type "(doi | pii | pmcpid | pmpid |  
sici | %art.id.type.int;)">

19.10.2010

```

<!--
===== --
>
<!ELEMENT PubmedArticleSet (PubmedArticle)+>
<!--
===== --
>
<!-- This is the top level element for PubMedArticle -->
<!ELEMENT PubmedArticle ((NCBIArticle | MedlineCitation), PubmedData?)>
<!--
===== --
>
<!ELEMENT PubmedData (History*, PublicationStatus, ArticleIdList)>
<!ELEMENT History (PubMedPubDate+)>
<!ELEMENT PubMedPubDate (%normal.date;)>
<!ATTLIST PubMedPubDate
    PubStatus %pub.status; #REQUIRED
>
<!ELEMENT PublicationStatus (#PCDATA)>
<!ELEMENT ArticleIdList (ArticleId+)>
<!ELEMENT ArticleId (#PCDATA)>
<!ATTLIST ArticleId
    IdType %art.id.type; "pubmed"
>
<!ELEMENT URL (#PCDATA)>
<!ATTLIST URL
    lang %iso.language.codes; #IMPLIED
    Type ( FullText | Summary | fulltext | summary) #IMPLIED
>
<!--
===== --
>

```