

Annotation Guideline for the Multi-Tagged Corpus (MTC)

The objective is to define an annotation schema that we can use to annotate entities in text in a way that is easy to understand and work with it and is complete enough to normalize different types of annotation. This document describes the way the annotations should be provided, points to a sample collection that has been prepared for this collaborative task and finally an example is provided as an appendix.

Annotation Guidelines Specification

Several points have been considered while defining the annotation schema:

- All the annotations will be based on XML since it is easier for the machine to process it and is human readable.
- Inline annotation is preferred since it is easier to handle and is more clear. Even though standoff annotation was proposed in the case of overlapping entities, an inline proposal by Olivier Bodenreider has been considered.
- A namespace is used to identify the concept in the original knowledge source.
- The exact boundaries of the entity are specified.
- Annotation of the largest-span of text is preferred. In the case of lung cancer, only *lung cancer* is annotated and not *lung* and/or *cancer*. This means that nested entities will not be annotated (but overlapping entities are still allowed).
- If there is no single concept for the entity in a given span of text, multiple concepts can be used (within a given knowledge source). The combined (post-coordinated) concepts, overlapping or not, could reflect the entity.
- Given an entity, if the system cannot decide on the identifier of the entity within a knowledge source all the identifiers are provided.
- Recovery of the original text.

In the remaining part of this section the annotation schema is defined. The annotation of entities is done using the *e element* that encloses the text where entity/entities can be found. The granularity of the annotation (exact boundaries, sentence, noun phrase, ...) provides different configurations; but the exact boundaries where the entity can be found is preferred. The entities are identified within the knowledge source identified using the *id attribute* in

the *e element*.

Since several sources can be used to identify the entities used in the annotation a distinctive namespace has to be used; the list of namespaces used by the different groups has to be provided with the description of the different sources they refer to.

In addition to the id from the original source, the semantic type of the concept has to be provided too. This semantic type will allow us to find the agreement on the semantic type even though different identifiers are provided by different participants; being relevant to build the super-annotation set.

The identifier of a given entity in a given datasource is composed of the namespace of the knowledge source (e.g., "UMLS"), the identifier of the entity in this source (e.g., "C0001403") and the semantic type. The namespace, the identifier and the semantic type are separated by the colon character (:) (e.g., "UMLS:C0001403"). In the case of multiple knowledge sources for the same *e element*, the pipe character is used (n1:id1:st1|n2:id2:st2). As an example, a piece of text that can be annotated with a Uniprot entity and a UMLS entity:

<e id="Uniprot: P01308:T028|UMLS:C1337112:T028">INS gene</e>

In the case that the entities are not constituted by a continuous stream of text or there are overlapping entities, we need to find a way to identify the pieces of text belonging to the entity. To do so, we need to divide the text in the annotation into smaller units (tokens) that are used in the annotation of the entity. The tokens are identified by the *w element*. There is no specific constraint about how the tokenization has to be performed. Each token will be identified with a token id and the id will depend on the system; the only constraint is that distinctively identifies the token.

After the entity identifier, specified above (namespace:id:semantic type), a colon indicates that there is a comma separated list of token identifiers. The following example illustrates this point:

<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". (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 the two concepts: "UMLS:C0205178:T079" for "acute" (token 1) and "UMLS:C0085437:T047" for "bacterial meningitis" (token 2). This example is illustrated below.

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

If the optional representation of tokens is not used, the two examples above can be represented more simply as follows. The first representation uses overlapping concepts, while the second does not.

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

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

Even though only the id attribute is mandatory, the usage of other attributes is not forbidden but a description has to be provided. Appendix A shows an example of annotated text.

The current version of the chosen resources has to be provided. Since it is a property of the annotated text, there is no reason to specify it at the annotation. This provides more freedom to the participants to explain in more detail the different processing of the resources. On the other hand the details have to be provided in the description of the work.

Minor issues:

- Annotation of text in the case that an entity is not assigned a concept: Span of text that is not attributed with an identified entity can be annotated with an attribute id value "none" for completeness reasons. This annotation is optional. In other words, if there is no explicit annotation for "unknown" spans of text in the document, this would still result into a valid annotation.

<e id="none">This is proteins</e><e id="uniprot:X:Y">BRCA1</e>

- Sentence level annotation. The element s was initially introduced for the annotation of a sentence. According to this specification it would also be possible to use the e element for the annotation of the whole sentence, where the annotation would deliver the precise location of the entities. Looks like a conflict, but since the current effort is focused on annotation of entities maybe we do not have to discuss it now. At the current state the test corpus will be delivered with pre-annotated sentences. This way we make sure that all discrepancies resulting from sentence detection are excluded from the start.

Sample Corpus Annotation

The idea of having the sample corpus is to provide an annotation of this corpus that will be used to compare the different groups and to improve this

annotation guideline. The corpus can be found at the following ftp site at the EBI <ftp://ftp.ebi.ac.uk/pub/software/textmining/corpora/immunology>

In this site you can find:

- docs*.xml.gz: Medline entries retrieved with the query "immunology" from PubMed
- immu.subset.xml.gz: subset of 150k documents from the immunology dataset
- smallSample.xml.gz: subset of 100 documents from the 150k documents

The annotation will be done on the specific sections of the document where the text can be found, this means the text found in the title (ArticleTitle field) and the abstract (AbstractText field).

Appendix A – Example of annotated text

Original document

```
<PubmedArticle><MedlineCitation Owner="NLM"
Status="MEDLINE"><PMID>1410221</PMID><DateCreated><Year>1992</Year><Month>11</Mo
nth><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</Yea
r><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><PageInfo><Page>367-
74</Page></PageInfo><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</Initial
s></Author></AuthorList><Language>eng</Language><PublicationTypeList><Publicatio
nType>Journal
Article</PublicationType></PublicationTypeList></Article><MedlineJournalInfo><Co
untry>ENGLAND</Country><MedlineTA>Public
Health</MedlineTA><NlmUniqueID>0376507</NlmUniqueID></MedlineJournalInfo><Chemic
alList><Chemical><RegistryNumber>0</RegistryNumber><NameOfSubstance>Diphtheria
Toxoid</NameOfSubstance></Chemical><Chemical><RegistryNumber>0</RegistryNumber><
NameOfSubstance>Pertussis
Vaccine</NameOfSubstance></Chemical></ChemicalList><CitationSubset>IM</CitationS
ubset><MeshHeadingList><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><DescriptorNa
me MajorTopicYN="N">Immunization
Schedule</DescriptorName></MeshHeading><MeshHeading><DescriptorName
MajorTopicYN="N">Infant</DescriptorName></MeshHeading><MeshHeading><DescriptorNa
me 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></Med
lineCitation><PubmedData><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></PubMedPubDate></History><PublicationStatus>ppublish</Publicat
ionStatus><ArticleIdList><ArticleId
```

IdType="pubmed">1410221</ArticleId><ArticleId
IdType="medline">93029182</ArticleId></ArticleIdList></PubMedData></PubMedArticle>

Annotated 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 <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> <e
id="umls:C0042196:T061:1|umls:C0019665:T170:2|umls:C0008059:T100:4"> <w
id="0">on the</w><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,<e
id="umls:C0205088:T079:1|umls:C0178602:T081:2,umls:C0012551:T109:4"><w
id="0">the</w> <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</e> <e id="umls:C1549113:T033:1|
umls:C1273803:T093:2"><w id="0">within a</w> <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</Initial
s></Author></AuthorList><Language>eng</Language><PublicationTypeList><Publicatio
nType>Journal
Article</PublicationType></PublicationTypeList></Article><MedlineJournalInfo><Co
untry>ENGLAND</Country><MedlineTA>Public
Health</MedlineTA><NlmUniqueID>0376507</NlmUniqueID></MedlineJournalInfo><Chemic
alList><Chemical><RegistryNumber>0</RegistryNumber><NameOfSubstance>Diphtheria
Toxoid</NameOfSubstance></Chemical><Chemical><RegistryNumber>0</RegistryNumber><
NameOfSubstance>Pertussis
Vaccine</NameOfSubstance></Chemical></ChemicalList><CitationSubset>IM</CitationS
ubset><MeshHeadingList><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><DescriptorNa
me MajorTopicYN="N">Immunization
Schedule</DescriptorName></MeshHeading><MeshHeading><DescriptorName
MajorTopicYN="N">Infant</DescriptorName></MeshHeading><MeshHeading><DescriptorNa
me 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></Med
lineCitation><PubmedData><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></PubMedPubDate></History><PublicationStatus>ppublish</Publicat
ionStatus><ArticleIdList><ArticleId
IdType="pubmed">1410221</ArticleId><ArticleId
IdType="medline">93029182</ArticleId></ArticleIdList></PubmedData></PubmedArticl
e>