Encoding Biomedical Resources in TEI: the Case of the GENIA Corpus

Tomaž Erjavec
Dept. of Intelligent Systems
Jožef Stefan Institute, Ljubljana

Jin-Dong Kim
Dept. of Information Science
University of Tokyo

Tomoko Ohta
CREST
Japan Science and
Technology Corporation

Yuka Tateisi
CREST
Japan Science and
Technology Corporation

Jun-ichi Tsujii
CREST JST &
Dept. of Information Science
University of Tokyo

Abstract

It is well known that standardising the annotation of language resources significantly raises their potential, as it enables re-use and spurs the development of common technologies. Despite the fact that increasingly complex linguistic information is being added to biomedical texts, no standard solutions have so far been proposed for their encoding. This paper describes a standardised XML tagset (DTD) for annotated biomedical corpora and other resources, which is based on the Text Encoding Initiative Guidelines P4, a general and parameterisable standard for encoding language resources. We ground the discussion in the encoding of the GENIA corpus, which currently contains 2,000 abstracts taken from the MEDLINE database, and has almost 100,000 hand-annotated terms marked for semantic class from the accompanying ontology. The paper introduces GENIA and TEI and implements a TEI parametrisation and conversion for the GENIA corpus. A number of aspects of biomedical language are discussed, such as complex tokenisation, prevalence of contractions and complex terms, and the linkage and encoding of ontologies.

1 Introduction

With the growing research on processing texts from the biomedical domain, the number of resources, esp. corpora, is increasing rapidly. Such corpora can be heavily annotated, e.g., with meta-data, words and part-of-speech tags, named entities, phrases, terms, concepts, translation equivalents, etc. Corpora are invaluable to the further development of technologies for utilising the information in biomedical texts, as they provide them with training and testing data. Given the value of such resources, it is important to ensure their reusability and increase their interchange potential — a step in this direction is developing common encodings for biomedical corpora.

Standardisation of resource encoding practices has now, for some time, been in the forefront of attention. Most of these advances are Web-driven, and include XML and related recommendations, such as XSLT, XML Schemas, XPointer, SAX, etc. The higher level standards, of meta-data (RDF) and ontologies (OWL) have been especially influential in encoding biomedical resources. However, there remains the question how to best encode the structure of the text themselves, how to mark-up added linguistic analyses, and how to implement linkages between the text and and further resources, such as lexica, thesauri and ontologies. As discussed in (Ide and Brew, 2000), in order to qualify as a "good" annotated corpus, its encoding should provide for reusabilty and extensibily.

In this paper we build on previous work (Erjavec et al., 2003) and show how to develop a standardised encoding for biomedical corpora. We base our discussion on the case of the GENIA corpus (Ohta et al., 2002), which is originaly encoded in GPML, the GENIA Project Markup Language, an XML DTD. We re-encode the corpus into a standardised annotation scheme, based on the Text Encoding Initiative Guidelines P4 (Sperberg-McQueen and Burnard, 2002), and specify a constructive mapping from the original DTD to the developed encoding via a XSLT transformation.

One of the motivations for such an re-encoding is that TEI is well-designed and widely accepted architecture, which has been often used for annotating language corpora, and by porting to it, GENIA, and other projects, can gain new insights into possible encoding practices and maybe make the corpus better suited for interchange. As the transformation to TEI is fully automatic, there is also no need to abandon the original markup format (in this case GPML), which, as it has been crafted specially for the corpus, provides a tighter encoding than can be possible with the more general TEI.

The paper thus proposes the creation of a practical annotation scheme for linguistically annotated (biomedical) corpora, the conversion to which is automatic and supports consistency checking and validation. The paper also serves as a guide to parametrising TEI and draws attention to certain aspects of biomedical corpora which are likely to face all that wish to process such texts.

The paper is structured as follows: Section 2 introduces the GENIA corpus; Section 3 introduces the TEI, gives some pros and cons of using it, and the method of parametrising TEI for particular projects; Section 4 discusses such a parametrisation for biomedical corpora and explains the conversion of the GENIA corpus to TEI; Section 5 discusses some challenging properties of biomedical text annotations; finally, Section 6 offers some conclusions and directions for further work.

2 The GENIA Corpus

The GENIA corpus (Ohta et al., 2002) is being developed in the scope of the GENIA project, which seeks to develop information extraction techniques for scientific texts using NLP technology. The corpus consists of semantically annotated published abstracts from the biomedical domain. The corpus is a collection of articles extracted from the on-line MEDLINE abstracts (U.S. National Center for Biotechnology Information, *http://www.ncbi.nlm.nih.gov/*, PubMed database). Since the focus of the corpus is on biological reactions concerning transcription factors in human blood cells, articles were selected that contain the MeSH terms *human*, *blood cell* and *transcription factor*.

As usual for the field, the articles are composed largely of structurally very complex technical terms, and are almost incomprehensible to a layperson. A typical heading e.g., reads *IL-2 gene expression and NF-kappa B activation through CD28 requires reactive oxygen production by 5-lipoxygenase.*

The main value of the GENIA corpus comes from its annotation: all the abstracts and their titles have been marked-up by two domain experts for biologically meaningful terms, and these terms have been semantically annotated with descriptors from the GENIA ontology.

The GENIA ontology is a taxonomy of, currently, 47 biologically relevant nominal categories, such as *body part*, *virus*, or *RNA domain or region*; the taxonomy has 35 terminal categories.

The terms of the corpus are semantically defined as those sentence constituents that can be categorised using the terminal categories from the ontology. Syntactically such constituents are quite varied: they include qualifiers and can be recursive.

The GENIA corpus is encoded in the Genia Project Markup Language. The GPML is an XML DTD (Kim et al., 2001) where each article contains its MEDLINE ID, title and abstract. The texts of the abstracts are segmented into sentences, and these contain the constituents with their semantic classification. The GENIA ontology is provided together with the GENIA corpus and is encoded in DAML+OIL (*http://www.daml.org/*), the standard XML-based ontology description language. This structure and its annotation will be further discussed below.

A suite of supporting tools has been developed or tuned for the GENIA corpus and GPML: the term annotation is performed with the XMLMind editor; an XPath-based concordancer has been developed for searching the corpus; and CSS stylesheets are available for browsing it.

At the time of writing, the latest version of the GENIA corpus is 3.01, which has been released in April 2003. It consists of 2,000 abstracts with over 400,000 words and more than 90,000 marked-up terms. This version has not yet been marked-up with tokens or PoS information, although an earlier version (Genia-V3.0p) has been. The GENIA corpus is available free of charge from the GENIA project homepage, at *http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/*.

3 The Text Encoding Initiative

The Text Encoding Initiative was established in 1987 as a systematised attempt to develop a fully general text encoding model and set of encoding conventions based upon it, suitable for processing and analysis of any type of text, in any language, and intended to serve the increasing range of existing (and potential) applications and uses. The TEI Guidelines for Electronic Text Encoding and Interchange were first published in April 1994 in two substantial green volumes, known as TEI P3. In May 1999, a revised edition of TEI P3 was produced, correcting several typographic and other errors. In December 2000 the TEI Consortium (*http://www.tei-c.org/*) was set up to maintain and develop the TEI standard. In 2002, the Consortium announced the availability of a major revision of TEI P3, the TEI P4 (Sperberg-McQueen and Burnard, 2002) the object of which is to provide equal support for XML and SGML applications using the TEI scheme. The revisions needed to make TEI P4 have been deliberately restricted to error correction only, with a view to ensuring that documents conforming to TEI P3 will not become illegal when processed with TEI P4. For GENIA, we are using the XML-compatible version of TEI P4.

In producing P4, many possibilities for other, more fundamental changes have been identified. With the establishment of the TEI Council, it became possible to agree on a programme of work to enhance and modify the Guidelines more fundamentally over the coming years. TEI P5 will be the next full revision of the Guidelines. The work on P5 has started, and the date of its appearance will likely be in 2004 and there are currently several TEI Working Groups addressing various parts of the Guidelines that need attention.

More than 80 projects spanning over 30 languages have so far made use of the TEI guidelines, producing diverse resources, e.g., text-critical editions of classical works. TEI has also been influential in corpus encoding, where the best known example is probably the British National Corpus. However, while the TEI has been extensively used for annotating PoS tagged corpora, it been less popular for encoding texts used by the the Information Retrieval/Extraction community; here, a number of other initiatives have taken the lead in encoding, say, ontologies or inter-document linking.

3.1 Pros and cons of using TEI

Why, if a corpus is already encoded in XML using a home-grown DTD, to re-encoded it in TEI at all? One reasons is certainly the validation aspect of the exercise: re-coding a corpus, or any other resource, reveals hidden (and in practice incorrect) assumptions about its structure. Re-coding to a standard recommendation also forces the corpus designers to face issues which might have been overlooked in the original design.

There are also other advantages of using TEI as the interchange format: (1) it is a wide-coverage, well-designed (modular and extensible), widely accepted and well-maintained architecture; (2) it provides extensive documentation, which comprises not only the Guidelines but also papers and documentation (best practices) of various projects; (3) it offers community support via the tei-l public discussion list; (4) various TEI-dedicated software already exists, and more is likely to become available; and (5) using it contributes to the adoption of open standards and recommendations.

However, using a very general recommendation which tries to cater for any possible situation brings with it also several disadvantages:

Tag abuse TEI might not have elements / attributes with the exact meaning we require. This results in a tendency to misuse tags for purposes they were not meant for; however, it is a case of individual judgement to decide whether to (slightly) abuse a tag, or to implement a local extension to add the attribute or element required.

Tag bloat Being a general purpose recommendation, TEI can — almost by definition — never be optimal for a specific application. Thus a custom developed DTD will be leaner, have less (redundant) tags and simpler content models.

TEI for humanities While the Guidelines cover a vast range of text types and annotations, they are maybe the least developed for "high level" NLP applications or have failed to keep abreast of "cutting-edge" initiatives. As will be seen, critical areas are the encoding of ontologies, of lexical databases and of feature structures.

3.2 Building the TEI DTD

The TEI Guidelines (Sperberg-McQueen and Burnard, 2002) consist of the formal part, which is a set of SGML/XML DTD fragments, and the documentation, which explains the rationale behind the elements available in these fragments, as well as giving overall information about the structure of the TEI.

The formal SGML/XML part of TEI comes as a set of DTD fragments or tagsets. A TEI DTD for a particular application is then constructed by selecting an appropriate combination of such tagsets. TEI distinguishes the following types of tagsets:

Core tagset : standard components of the TEI main DTD in all its forms; these are always included without any special action by the encoder.

Base tagsets : basic building blocks for specific text types; exactly one base must be selected by the encoder, unless one of the combined bases is used.

Additional tagsets : extra tags useful for particular purposes. All additional tagsets are compatible with all bases and with each other; an encoder may therefore add them to the selected base in any combination desired.

User defined tagsets : these extra tags give the possibility of extending and overriding the definitions provided in the TEI tagset. Furthermore, they give the option of explicitly including or ignoring (disallowing) each particular element licensed by the chosen base and additional tagsets.

```
<!DOCTYPE teiCorpus.2 SYSTEM
 "http://www.tei-c.org/P4X/DTD/tei2.dtd"
 [<!ENTITY % TEI.XML          "INCLUDE">
  <!ENTITY % TEI.prose        "INCLUDE">
  <!ENTITY % TEI.linking      "INCLUDE">
  <!ENTITY % TEI.analysis     "INCLUDE">
  <!ENTITY % TEI.corpus       "INCLUDE">
  <!ENTITY % TEI.extensions.ent SYSTEM
           'geniaex.ent'>
  <!ENTITY % TEI.extensions.dtd SYSTEM
           'geniaex.dtd'>
 ]>
```

Figure 1: The XML TEI prolog for GENIA

While a project-particular XML DTD can be constructed by including and ignoring the TEI DTD fragments directly (as exemplified in Figure 1), it is also possible to build — for easier processing — a one-file DTD with the help of the on-line TEI Pizza Chef service, available from the TEI web site.

4 Parametrising TEI for biomedical corpora

In previous work (Erjavec et al., 2003) we have already proposed a TEI parametrisation of GENIA which was quite broad in its scope. Because a number of tagsets could prove useful in the long term this parametrisation collected not only those that we considered necessary for the current version of GENIA, but also some that might prove of service in the future. Furthermore, we supported the encoding of both version 2.1 and 3.0 of the corpus. The resulting DTD was thus very generous in what kinds of data it caters for. To focus the discussion we, in the current paper, only address tagset that are immediately relevant to annotating biomedical texts. In Figure 1 we define the XML DTD that can be used for encoding biomedical resources, and that we used for GENIA V3.01. The XML prolog given in this Figure defines that $\langle$*teiCorpus.2*$\rangle$ is the root element of the corpus, that the external DTD resides at the given URL belonging to the TEI Consortium, and that a number of TEI modules, detailed below, are being used to parametrise the TEI to arrive at our particular DTD.

4.1 TEI.XML

TEI P4 allows both standard SGML and XML encodings. Including the TEI.XML option indicates that the target DTD is to be expressed in XML.

4.2 TEI.prose

The base tagset does not declare many elements but rather inherits all of the TEI core, which includes the TEI header, and text elements. A TEI document will typically have as its root element ⟨*TEI.2*⟩ which is composed of the ⟨*teiHeader*⟩, followed by the ⟨*text*⟩; c.f. right hand side of Figure 2, but note that the root element from the TEI.corpus module is used for the complete corpus.

The TEI header describes an encoded work so that the text (corpus) itself, its source, its encoding, and its revisions are all thoroughly documented.

TEI.prose also contains elements and attributes for describing text structure, e.g. ⟨*div*⟩ for text division, ⟨*p*⟩ for paragraph, ⟨*head*⟩ for text header, etc. The tagset is therefore useful for encoding the gross structure of the corpus texts; for an illustration again see Figure 2.

4.3 TEI.linking

This additional tagset provides mechanisms for linking, segmentation, and alignment. The elements provided here enable links to be made e.g., between the articles and their source URLs, or between concepts and their hypernyms.

It should be noted that while the TEI treatment of external pointers had been very influential, it was overtaken and made obsolete by newer recommendations. However, the TEI does have a Working Group on Stand-Off Markup, XLink and XPointer, which should produce new TEI encoding recommendations for this area in 2003.

4.4 TEI.analysis

This additional tagset is used for associating simple linguistic analyses and interpretations with text elements. It can be used to annotate words, ⟨*w*⟩, clauses, ⟨*cl*⟩, and sentences, ⟨*s*⟩ with dedicated tags, as well as arbitrary and possibly nested segments with the ⟨*seg*⟩. Such elements can be, via attributes, associated with their analyses. This tagset has proved very popular for PoS-annotated corpora; for an illustration see Figure 3.

4.5 TEI.corpus

This additional tagset introduces a new root element, ⟨*teiCorpus.2*⟩, which comprises a (corpus) header and a series of ⟨*TEI.2*⟩ elements. The TEI.corpus tagset also extends the certain header elements to provide more detailed descriptions of the corpus material.

4.6 TEI.extensions.ent

The file gives, for each element sanctioned by the chosen modules, whether we include or ignore it in our parametrisation. While this is not strictly necessary (without any such specification, all the elements would be included) we thought it wise to constrain the content models somewhat, to reduce the bewildering variety of choices that the TEI otherwise offers. Also, such an entity extension file gives the complete list of all the TEI elements that are allowed (and disallowed) in GENIA, which might prove useful for documentation purposes.

4.7 TEI.extensions.dtd

This file specifies the changes we have made to TEI elements. We have e.g., added the *url* attribute to ⟨*xptr*⟩ and ⟨*xref*⟩ and tagging attributes to word and punctuation elements.

4.8 Conversion of GPML to TEI

Because the source format of GENIA will remain the simpler GPML, it is imperative to have an automatic procedure for converting to the TEI interchange format. The translation process takes advantage of the fact that both the input and output are encoded in XML, which makes it possible to use the XSL Transformation Language, XSLT that defines a standard declarative specification of transformations between XML documents. There also exist a number of free XSLT processors; we used Daniel Veillard's `xsltproc`.

The transformation is written as a XSLT stylesheet, which makes reference to two documents: the GENIA ontology in TEI and the template for the corpus header. The stylesheet then resolves the GPML encoded corpus into TEI. The translation of the corpus is thus fully automatic, except for the taxonomy, which was translated by hand.

Figure 2 illustrates the top level structure of the corpus, and how it differs between the GPML and

TEI encodings. The most noticeable difference is, apart from the renaming of elements, the addition of headers to the corpus and texts. In the GENIA ⟨*teiHeader*⟩ we give e.g., the name, address, availability, sampling description, and, for each abstract's ⟨*sourceDesc*⟩, two ⟨*xptr*⟩s: the first gives the URL of the HTML article in the MEDLINE database, while the second is the URL of the article in the original XML. It should be noted that we use a locally defined *url* attribute for specifying the value of the pointer.

5 Characteristics of biomedical texts

In this section we review some challenges that biomedical texts present to the processing and encoding of linguistic information, and the manner of their encoding in our DTD.

5.1 Tokens

Tokenisation, i.e., the identification of words and punctuation marks, is the lowest level of linguistic analysis, yet is, in spite (or because) of this of considerable importance. As all other levels of linguistic markup make direct or direct reference to the token stream of the text, so if this is incorrect, errors will propagate to all other annotations.

It is also interesting to note that current annotation practice is more and more leaning toward standoff markup, i.e., annotations that are separated from the primary data (text) and make reference to it only via pointers. However, it is beneficial to have some markup in the primary data to which it is possible to refer, and this markup is, almost exclusivelly, that of tokens; see e.g., (Freese et al., 2003).

Version V1.1 of GENIA has been also annotated with LTG tools (Grover et al., 2002). In short, the corpus is tokenised, and then part-of-speech tagged with two taggers, each one using a different tagset, and the nouns and verbs lemmatised. Additionally, the deverbal nominalisations are assigned their verbal stems.

The conversion to TEI is also able to handle this additional markup, by using the TEI.analysis module. The word and punctuation tokens are encoded as ⟨*w*⟩ and ⟨*c*⟩ elements respectively, which are further marked with *type* and *lemma* and the locally defined *c1*, *c2* and *vstem*. An example of such markup is given in Figure 3.

```
<s>
<w c1="DT" c2="DB">All</w>
<c type="HYPH" c1=":" c2="-">-</c>
<w c1="VBZ" c2="JJ">trans</w>
<w c1="JJ" c2="JJ">retinoic</w>
<w lemma="acid" c1="NN" c2="NN1">acid</w>
<c type="BR" c1="(" c2="(">(</c>
<w lemma="Ra" c1="NN" c2="NP1">RA</w>
<c type="BR" c1=")" c2=")">)</c>
<w lemma="be" c1="VBZ" c2="VBZ">is</w>
<w c1="DT" c2="AT1">an</w>
<w c1="JJ" c2="JJ">important</w>
...
```

Figure 3: TEI encoding of annotated tokens

Given the high density of technical terms, biomedical texts are rife with various types of contractions, such as abbreviations, acronyms, prefixes, etc. As seen already in Figure 3, one of the more problematic apects of tokenisaton are parentheses. Almost all tokenisers (e.g., the LT one, or the UPENN tokeniser) take these as separate tokens, but many are in biomedical texts parts of terms. So, out of almost 35,000 distinct terms that have been marked up in the GENIA corpus, over 1,700 contain parentheses. Some examples: *(+)-pentazocine*, *(3H)-E2 binding*, *(gamma(c))-like molecule*.

Correct tokenisation of the biomedical texts is thus a challenging tasks, and it is fair to say that, from a linguistic processing perspective, complex tokenisation is one of the defining characteristics of such corpora.

5.2 Terms

Annotation of terms is a prerequisite for meaningful processing of biomedical texts, yet it is often difficult to decide what constitutes a term in a text, and how to abstract away from local variations. Biomedical texts are largerly (one could almost say excusivelly) composed of terms, and, as mentioned, this brings with it complex abbreviatory mechanisms.

Even though TEI offers a ⟨*term*⟩ element, we chose, in line with the original GPML encoding, to rather use the TEI.analysis clause (⟨*cl*⟩) element to encode terms. In GENIA, the terms have been hand-annotated, and marked up with concepts from the GENIA ontology; this was also the defining factor of term-hood, namely that the term could be linked

```
<!DOCTYPE set SYSTEM "gpml.dtd">
<set>
 <article>
  <articleinfo><bibliomisc>
     *MEDLINE_ID*
   </bibliomisc></articleinfo>
  <title>
   *Title_of_article*
  </title>
  <abstract>
   *Abstract_of_article*
  </abstract>
 </article>
 *More_articles*
</set>
```

```
<!DOCTYPE teiCorpus.2 SYSTEM "genia-tei.dtd">
 <TEIcorpus.2>
  <teiHeader type="corpus">
   *Corpus_header*</teiHeader>
   <TEI.2 id="*MEDLINE_ID*">
     <teiHeader type="text">
      *Article_header*</teiHeader>
     <text><body>
      <div type="abstract">
       <head>*Title_of_article*</head>
       <p>*Abstract_of_article*</p>
      </div>
     </body></text></TEI.2>
   *More_articles*
   </TEIcorpus.2>
```

Figure 2: The GPML and TEI structure of the corpus

to a terminal concept of the GENIA ontology.

In spite of the simple semantic definition, the syntactic structure of the terms in the corpus varies dramatically. Biomedical terms are in some ways similar to named entities (names of people, organizations, etc.) but from the linguistic perspective, they are different in that named entities are mostly proper nouns, while terms mostly contain common nouns, and the two differ in their syntactic properties. Terms in the corpus can also be nested, where complex terms are composed out of simpler ones, e.g., ⟨*cl*⟩⟨*cl*⟩*IL-2 gene*⟨/*cl*⟩ *transcription*⟨/*cl*⟩.

This nesting, and the reference to ontology concepts is often far from simple, as (partial) terms can appear in coordinated clauses involving ellipsis. For example, "*CD2 and CD 25 receptors*" refers to two terms, *CD2 receptors* and *CD25 receptors*, but only the latter actually appears in the text.

In such cases by parsing the coordination all the terms can be identified and annotated; the TEI encoding achieves this by specifyng the propositional formula involving the participating concepts in the *function* attribute; for example, ⟨*cl function="(AND G.tissue G.tissue)" ana="G.tissue"*⟩⟨*cl*⟩*normal*⟨/*cl*⟩ *and* ⟨*cl*⟩*hypopigmented*⟨/*cl*⟩ ⟨*cl*⟩*skin samples*⟨/*cl*⟩⟨/*cl*⟩.

The *ana* attribute encodes the IDREF of the concept; currently, only same valued concepts are either conjoined or disjoined.

The number of ⟨*cl*⟩ elements in the GENIA corpus is 96,582, among which 89,682 are simple terms and 1,583 are nested terms that are contain 3,431 terms. 5,137 terms do not yet have the *ana* attribute for concept identification, so the total number of ontology-linked terms is 93,293.

5.3 Ontologies

One of the more interesting questions in recoding GENIA in TEI was how to encode the ontology. The ontology is in GENIA GPML encoded in a separate document, conforming to the OIL+DAML specification. This, inter alia, means that that XML file heavily relies on XML Namespaces and the RDF recommendation. An illustrative fragment is given on the left side of Figure 4.

Currently the GENIA ontology has a simple tree-like structure, i.e., it corresponds to a taxonomy, so we translated it to the TEI ⟨*taxonomy*⟩ element, which is contained in the ⟨*classDecl*⟩ of the header ⟨*encodingDesc*⟩. The TEI defines this element as "[the classification declaration] contains one or more taxonomies defining any classificatory codes used elsewhere in the text", i.e., is exactly suited for our purposes.

There are quite substantial differences between the two encodings: the DAML+OIL models class inclusion with links, while the TEI does it as XML element inclusion. This is certainly the simpler and more robust solution, but requires that the ontology is a taxonomy, i.e., tree structured. The second difference is in the status of the identifiers: in DAML+OIL they are general `#CDATA` links, which need a separate (XLink/XPointer) mechanisms for their resolution. In TEI they are XML `ID` attributes,

```
<daml:Class rdf:ID="source"></daml:Class>
<daml:Class rdf:ID="natural">
 <rdfs:subClassOf rdf:resource="#source"/>
</daml:Class>
<daml:Class rdf:ID="organism">
 <rdfs:subClassOf rdf:resource="#natural"/>
</daml:Class>
<daml:Class rdf:ID="multi_cell">
 <rdfs:subClassOf rdf:resource="#organism"/>
</daml:Class>
...
```

```
<taxonomy id="G.taxonomy">
 <category id="G.source">
  <catDesc>biological source</catDesc>
  <category id="G.natural">
   <catDesc>natural</catDesc>
   <category id="G.organism">
    <catDesc>organism</catDesc>
    <category id="G.multi_cell">
     <catDesc>multi-cellular</catDesc>
    </category>
...
```

Figure 4: The GENIA DAML+OIL and TEI ontology

and can rely on the XML parser to resolve them. While this is a simpler solution, it does support document-internal reference only.

6 Conclusions

The paper proposed an XML paramterisation of TEI P4 developed for linguistically annotated biomedical corpora, and applied it to the GENIA corpus. The conversion from the Genia Project Markup Language to this encoding has been implemented in XSLT and both the TEI-conformant parametrisation (TEI extension file and one-file DTD) and the XSLT stylesheets are, together with a report documenting them, available at *http://nl.ijs.si/et/genia/*, while the GENIA corpus is freely available from *http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/*.

The paper gave a survey of the TEI modules that can be useful for encoding a wide variety of linguistically annotated corpora. This contribution, it is hoped, can thus serve as a blueprint for parametrising TEI for diverse corpus resources.

Further work involves the inclusion of other knowledge sources into the corpus, say of Medical Subject Headings (MeSH), Unified Medical Language System (UMLS), International Classification of Disease (ICD), etc. The place of these annotations in the corpus will have to be considered, and their linking to the existing information determined.

References

Tomaž Erjavec, Jin-Dong Kim, Tomoko Ohta, Yuka Tateisi, and Jun ichi Tsujii. 2003. Stretching the TEI: Converting the GENIA corpus. In *Proceedings of the EACL-03 Workshop on Linguistically Interpreted Corpora (LINC-03)*, pages 117–124, Budapest. ACL.

Marion Freese, Ulrich Heid, and Martin Emele. 2003. Enhancing XCES to XCOMFORT: An Extensible Modular Architecture for Manipulation of Text Resources. In *Proceedings of the EACL-03 Workshop on Language Technology and the Semantic Web: 3rd Workshop on NLP and XML (NLPXML-2003)*, pages 33–40, Budapest. ACL.

Claire Grover, Ewan Klein, Alex Lascarides, and Maria Lapata. 2002. XML-based NLP Tools for Analysing and Annotating Medical Language. In *2nd Workshop on NLP and XML (CoLing Workshop NLPXML-2002)*. http://www.ltg.ed.ac.uk/software/ttt/.

Nancy Ide and Chris Brew. 2000. Requrements, Tools and Architectures for Annotated Corpora. In *Proceedings of Data Architectures and Software Support for Large Corpora*, pages 1–5, Budapest. ELRA.

Jin-Dong Kim, Tomoko Ohta, and Jun-ichi Tsujii. 2001. XML-based Linguistic Annotation of Corpus. In *Proceedings of the first NLP and XML Workshop*, pages 44–53.

Tomoko Ohta, Yuka Tateisi, and Jin-Dong Kim. 2002. The GENIA Corpus: an Annotated Research Abstract Corpus in Molecular Biology Domain. In *Proceedings of the Human Language Technology Conference*, page To appear.

C. M. Sperberg-McQueen and Lou Burnard, editors. 2002. *Guidelines for Electronic Text Encoding and Interchange, The XML Version of the TEI Guidelines*. The TEI Consortium. http://www.tei-c.org/.