

Texttechnological Standards

An Overview

GLDV 2007

April 12th 2007

Maik Stührenberg
Bielefeld University

Overview of this talk

- Standards – do we need them?
- Standards for annotating large text corpora
- Metadata – data about data
- Outline of a harmonised XML Framework
- Conclusion and outlook

Standards – do we need them?

- The term standard is used for normative specifications as well as for de-facto-standards
- In this talk we will refer mainly to specifications that are widely used and therefore accepted as de-facto-standards

Standards – do we need them?

- Although this talk is about standards in the fields of linguistic data...
- ...take a moment to imagine a world without standards

Standards – do we need them?

Imagine a world...

- ...where you could only buy paper sheets from the distributor of your printer – because there is no A4 or letter size
- ...where you could only buy shoes from your very special dealer – because shoe sizes aren't standardised at all

Standards – do we need them?

But on the other side...

- ...there are too many standards (just imagine the several different shoe size standards in the world)

Standards – do we need them?

As a quintessence:

- Standards are useful for our daily life
- There is a variety of standards and sometimes this makes it hard to choose the right one for one's needs

Standards for annotating large text corpora

This talk is about choosing the right annotation format
and

about showing you an alternative XML annotation representation

Standards for annotating large text corpora

- When we're talking about standards for large textual corpora we will exclusively talk about XML markup languages for annotating unimodal corpora, i.e.
- DocBook
- TEI Guidelines
- CES and XCES

DocBook

- One of the most used specifications in annotating texts
- Powerful in structuring various types of documents:
 - help files
 - books
 - articles
 - and many more
- XSLT stylesheets for transforming DocBook to HTML, PostScript or PDF are available

DocBook

- OASIS standard
- Available as SGML and XML DTD, XML W3C Schema and RELAX NG Schema
- About 400 elements
- Elements of other namespaces can be included as well

DocBook

Structuring DocBook documents:

- Set, Book and Article as main classes
- Divisions divide books into smaller parts
- Components divide books or divisions into chapters
- Sections subdivide components
- Block level elements
- Inline elements

DocBook

DocBook and corpus annotation:

- DocBook lacks specific elements for storing whole corpora
- No elements for linguistic phenomena are provided

TEI Guidelines

- Started as SGML application in 1987 (just like DocBook)
- Maintained by the TEI Consortium
- Family of markup languages
- Available as XML DTD, XML W3C Schema, RELAX NG Schema

TEI Guidelines

Three building blocks:

- Core tag set (containing the TEI header and elements available in all documents)
- Base tag set according to specific text types
- Additional tag sets for particular purposes

TEI Guidelines

Base tag sets:

- TEI.prose
- TEI.verse
- TEI.drama
- TEI.spoken
- TEI.dictionaries
- TEI.terminology
- TEI.general
- TEI.mixed

TEI Guidelines

Additional tag sets for:

- hypertext
- linking
- tables
- figures
- linguistic analyses
- language corpora
- and some more

TEI Guidelines

TEI Guidelines and corpus annotation:

- Elements for several text divisions are available, going down to single character markup
- Elements for speech phenomena are provided in the TEI.spoken base tag set
- Overlapping and segmentation is supported as well

CES and XCES

- Corpus Encoding Standard (CES) developed as SGML application, XCES as XML application
- TEI compliant
- Available as XML W3C Schema
- Designed especially for annotation of large text corpora
- Primary data and annotation are separated, CES introduced standoff annotation

CES and XCES

(X)CES and corpus annotation:

- In principle, XCES is the best suitable annotation format for corpus data
- But the supplied W3C schema files are not valid and the development of the specification has been ceased

Other specifications

- The Linguistic Annotation Framework (LAF) which is developed by ISO/IEC TC37 SC4
- ...

Metadata – data about data

- Metadata included in the before mentioned specifications (built-in metadata)
- Independent Metadata specifications

Independent Metadata specifications

- Dublin Core
- OLAC
- IMDI

Dublin Core

- ISO standard developed by the Dublin Core Metadata Initiative
- Dublin Core Metadata Element Set containing 15 metadata elements which can be divided into the following sections:
 - Content
 - Intellectual Property
 - Instantiation
- DCMI Metadata Terms includes additional elements and element refinements and encoding schemes

Dublin Core

- Available in several formats, DC can be used in (X)HTML, XML, RDF/XML
- However, DC is very general and lacks specific linguistic metadata elements

OLAC

- OLAC Metadata set, developed by the Open Language Archive Community
- Based on Dublin Core
- Provides greater precision in resource description for the linguistic community by refinement or encoding schemes

IMDI

- International Standard for Language Engineering (ISLE) Meta Data Initiative (IMDI)
- IMDI specifies metadata for multi-lingual and multimodal corpora
- Separation between catalogue and session metadata
- Metadata elements for lexicon description and a vocabulary taxonomy are available as well

Outline of a harmonised XML framework

Ingredients:

- A modular approach
- The best components of the before mentioned specifications, i.e.
 - A corrected version of XCES (including the current TEI as base)
 - IMDI metadata
- And a little bit more...

Outline of a harmonised XML framework

Imagine the following modular framework...

Outline of a harmonised XML framework

A root layer

```
<corpus>  
  <corpusdata id="text1">  
  </corpusdata>  
</corpus>
```

Outline of a harmonised XML framework

A base layer containing character data, token or time based events in case of multimodal corpora

```
<corpus xmlns:cdata="http://www.text-technology.de/sekimo/cdata">
  <corpusdata id="text1">
    <cdata:text startpos="0" endpos="18">This is a sentence.</cdata:text>
    <cdata:whitespace startpos="4" endpos="4" charref="0020" />
    <cdata:whitespace startpos="7" endpos="7" charref="0020" />
    <cdata:whitespace startpos="9" endpos="9" charref="0020" />
    <cdata:punctuation startpos="18" endpos="18" charref="002E" />
  </corpusdata>
</corpus>
```

Outline of a harmonised XML framework

Additional layers can be linked to the text/token/time base either via the character position, the token id or the event id

```
<corpus xmlns:cdata="http://www.text-technology.de/sekimo/cdata"
xmlns:syll="http://www.text-technology.de/sekimo/syll">
  <corpusdata id="text1">
    <cdata:text startpos="0" endpos="18">This is a sentence.</cdata:text>
    <cdata:whitespace startpos="4" endpos="4" charref="0020" />
    <cdata:whitespace startpos="7" endpos="7" charref="0020" />
    <cdata:whitespace startpos="9" endpos="9" charref="0020" />
    <cdata:punctuation startpos="18" endpos="18" charref="002E" />
    <syll:syll startpos="0" endpos="18">
      <syll:s startpos="0" endpos="3" />
      <syll:s startpos="5" endpos="6" />
      <syll:s startpos="8" endpos="8" />
      <syll:s startpos="10" endpos="12" />
      <syll:s startpos="13" endpos="17" />
    </syll:syll>
  </corpusdata>
</corpus>
```

Outline of a harmonised XML framework

You choose...

- ...the kind of segmentation
- ...how general or precise the annotation is

- In fact, you could choose DocBook for structuring means and XCES for linguistic annotation
- In fact, you could use your own existing annotation format and enrich it with another annotation format

Outline of a harmonised XML framework

The whole corpus data is stored in a native XML database, i.e. eXist

Why a native XML database?

- Easy way to store stand alone annotations correlating to the same primary data
- XQuery 1.0 and XPath 2.0 for retrieval and updates
- And a straight-forward solution when XPath 2.0 Full Text is available

Conclusion and outlook

- If you'd like to start a new annotation project which deals with large text corpora, choose an existing specification – preferable a corrected version of XCES or the current TEI
- If you have to save concurrent markup you should wait until the architecture described in this talk or the LAF is available

Questions or Comments?

maik.stuehrenberg@uni-bielefeld.de