# The MLCD Overlap Corpus (MOC)

**Huitfeldt, Claus**
Claus.Huitfeldt@uib.no
Department of Philosophy, University of Bergen

**Sperberg-McQueen, C. M.**
cmsmcq@blackmesatech.com
Black Mesa Technologies LLC, USA

**Marcoux, Yves**
ymarcoux@gmail.com
Université de Montréal, Canada

---

For some time, theorists and practitioners of descriptive markup have been aware that the strict hierarchical organization of elements provided by SGML and XML represents a potentially problematic abstraction. The nesting structures of SGML and XML capture an important property of real texts and represent a successful combination of expressive power and tractability. But not all textual phenomena appear in properly nested form, and for more than twenty years students of markup have been exploring methods of recording overlapping (non-hierarchical) structures. Useful surveys include (Barnard et al. 1995), (DeRose 2004), and (Witt et al. 2005).

Some approaches to the overlap problem take the form of non-SGML, non-XML syntaxes and non-tree-like data structures. One example is offered by the TexMecs syntax and Goddag data structures proposed by the project Markup Languages for Complex Documents (MLCD) based at the University of Bergen. Another is the Layered Markup and Annotation Language (LMNL). A third is the so-called multi-colored trees defined by (Jagadish et al. 2004).

Other approaches exploit the optional *concurrent markup* feature of SGML (Sperberg-McQueen and Huitfeldt 1998), or apply it, with suitable modifications, to XML (Hilbert et al. 2005).

But by far the largest number of published approaches to problems of overlapping markup involve the use of SGML and XML themselves to record the information. They exploit the semantic openness of SGML and XML to supply non-hierarchical interpretation of what are often thought to be inescapably hierarchical notations.

The SGML/XML-based approaches to overlap fall, roughly, into three groups: milestones, fragmentation-and-reconstitution, and stand-off annotation. Milestones (described as early as (Barnard et al. 1988), and used in (Sperberg-McQueen and Burnard 1990) and later versions of the TEI *Guidelines*) use empty elements to mark the boundaries of regions which cannot be marked simply as elements because they overlap the boundaries of other elements. More recently, approaches to milestone markup have been generalized in the Trojan Horse and CLIX markup idioms (DeRose 2004).

Fragmentation is the technique of dividing a logical unit which overlaps other units into several smaller units, which do not; the consuming application can then re-aggregate the fragments.

Stand-off annotation addresses the overlap problem by removing the markup from the main data stream of the document, at the same time adding pointers back into the base data. Many language corpora use forms of stand-off markup (e.g. (Carletta et al. 2005), (Witt et al. 2005), (Stührenberg and Goecke 2008)).

For all the variety of methods and proposals for handling overlap, there is remarkably little consensus on the best approach. Even systematic comparisons are scarce, although several of the surveys provide at least a broad categorization of methods. Partly this reflects a pragmatic issue (many methods used in production systems are devised for use by specific projects, which do not wish to engage in a systematic comparison of interest to markup theorists, but to get on with their discipline-specific work); partly it reflects a difficulty in comparing different schemes point to point, owing to the scattered and informal nature of the documentation.

And finally, despite the work of the last twenty years we still have only an incomplete understanding of the different structural and semantic forms of overlapping structure, and the implications for markup practice of different forms of overlap. The pervasive but unsystematic overlap of verse and dramatic

structure in verse drama, or of formal and physical structure in any printed book, seems to present one kind of phenomenon. The occasional but richly significant overlap of structures characteristic of enjambement in verse may appear, on the other hand, to be of a different kind. Is it?

The MLCD Overlap Corpus (MOC) is intended to make it easier to compare different methods of handling overlap, not just on theoretical or abstract grounds, but in terms of concrete examples from real and constructed texts. The essential idea of the corpus is to make available a single body of material, ranging from compact examples to full texts of novel or five-act-play length, tagged for the same information, using a variety of overlap notations.

Consider the following simple example (from (Hilbert et al. 2005)) of a discourse situation in which the utterance structure overlaps with the syntactic structure:

**Peter**: Hey, Paul! Would you give me

**Paul**: the hammer?

(Hilbert et al. 2005) give the following representation of this example in the notation now known as XConcur (then MLX).

```
<!DOCTYPE (1)div SYSTEM "tei/dtd/teispok2.dtd">
<!DOCTYPE (2)text SYSTEM "tei/dtd/teiana2.dtd">
   <(1)div type="dialog" org="uniform">
       <(2)text>
          <(1)u who="Peter">
             <(2)s>Hey Paul!</(2)s>
             <(2)s>Would you give me
          </(1)u>
          <(1)u who="Paul">
             the hammer?</(2)s>
          </(1)u>
       </(2)text>
    </(1)div>
```

Using the CONCUR feature of SGML, a very similar representation can be given (elided here for space reasons). It might be represented in TexMecs this way:

```
<div type="dialog" org="uniform">
  <u who="Peter">
      <s>Hey Paul!</s>
      <s sID="s2"/>Would you give me
  </u>
  <u who="Paul">
      the hammer?<s eID="s2">
  </u>
</div>
```

The goal of MOC is to make examples available in a broad variety of notations, as well as those just given:

- various forms of TEI markup, using different TEI mechanisms (*next* and *prev* attributes, the *part* attribute, virtual elements, stand-off markup using feature structures, etc.)
- TexMecs (Huitfeldt and Sperberg-McQueen 2003)
- XStandoff (Stührenberg and Jettka 2009)
- Multix (Chatti et al., 2007)
- Sekimo General Format (SGF) (Stührenberg and Goecke 2008)
- Nite (Carletta et al. 2005)
- Earmark (Di Iorio et al. 2009)

There will be three sets of data:

- twenty or more 'toy' examples like the one just given, typically just a few lines in length. Most of the toy examples will be drawn from existing literature on overlap; almost all of them will be constructed texts, though some will be very short extracts from literary or other natural texts.
- ten or more 'short' examples, typically corresponding to a few pages of printed material, mostly extracts from natural texts.
- five or more 'long' examples, full-length natural texts. We will draw these partly from an existing collection of literary texts used as a test bed for full-text software and partly from existing language corpora.

The toy examples will be tagged manually in the various notations selected. The short examples will be tagged using semi-automated processes (i.e. partly by hand and partly automatically), and checked carefully for correctness. The long examples will be tagged using mostly automated processes, and checked carefully for correctness.

Since the purpose of MOC is to illuminate problems connected with overlap and with existing proposals for handling it, there will be no attempt to make the selection of texts representative of any particular natural language community. The relevant population is not a particular set of natural-language users, but the set of people who work with natural-language texts for various purposes. In such a

small corpus, we cannot and do not hope for statistical representativeness, but only for an illuminating variety of examples. Accordingly, we will seek to include examples illustrative of problems encountered in:

- literary and lexicological study
- metrical study
- language corpora (discourse analysis, syntax, prosody, ...)
- change markup and multi-versioned texts
- historical-critical editions
- analytic bibliography
- historical annotation

Apart from simply illustrating the ways in which different notations represent the same information, MOC should provide sample test data useful for a variety of tasks and studies:

- development of automatic translation among notations (the existing samples of the target notation serve as comparison points for the results achieved by the automatic translator)
- development of software intended to handle any of the notations represented
- construction of domain-appropriate queries against the various notations (does notation N1 make it easier to construct suitable queries than notation N2?)
- comparative measures of markup complexity
- analysis of different kinds and forms of overlap: do structural patterns vary with different kinds of markup? Do the domain-specific implications of overlap (and thus the domain-oriented requirements for manipulating the data) vary?
- development of tools for automatic extraction of formalized representations of the meaning of markup

Performance comparisons are notably missing from this list; MOC will be too small to provide performance measurements relevant to searches across typical modern collections in the gigabyte size range. (On the other hand, the long samples may be useful for at least preliminary performance comparisons and preparation for more large-scale testing.)

At the time this abstract is prepared, the first version of MOC is expected to be partially completed before the DH 2010 conference; the presentation will include an account of the work to date, problems encountered, and a forecast of the work remaining before completion of the corpus.

Follow-on work includes experimentation with existing full-text indexing and query systems to test the different characteristics of different markup styles on query formulation and retrieval time; we also expect to work on automated translations among various notations.

---

## References

**Barnard, D., Hayter, R., Karababa, M., Logan, G. and McFadden, J.** (1988). 'SGML Markup for Literary Texts'. *SGML Markup for Literary Texts.* **22**: 265-276.

**Barnard, D., Burnard, L., Gaspart, J. P., Price, L. A., Sperberg-McQueen, C. M. and Varile, G. B.** (1995). 'Hierarchical encoding of text: Technical problems and SGML solutions'. *Computers and the Humanities.* **29**: 211-231.

**Carletta, J., Evert, S., Heid, U. and Kilgour, J.** (2005). 'The NITE XML Toolkit: data model and query'. *Language Resources and Evaluation.* **39(4)**: 313-334. doi:10.1007/s10579-006-9001-9.

**Chatti, N., Kaouk, S., Calabretto, S. and Pinon, J. M.** (2007). 'MultiX: an XML-based formalism to encode multi-structured documents'. *Proceedings of Extreme Markup Languages 2007.* Montréal (Canada), Aug. 2007. http://conferences.idealliance.org/extreme/html/2007/Chatti01/EML2007Chatti01.html.

**DeRose, S. J.** (2004). 'Markup overlap: A review and a horse'. *Proceedings of Extreme Markup Languages 2004.* Montréal (Canada), Aug. 2004. http://conferences.idealliance.org/extreme/html/2004/DeRose01/EML2004DeRose01.html.

**Di Iorio, A., Peroni, S. and Vitali, F.** (2009). 'Towards markup support for

full GODDAGs and beyond: the EARMARK approach'. *Proceedings of Balisage: The Markup Conference 2009*. Montréal (Canada), August 11-14, 2009. http://www.balisage.net/Proceedings/vol3/html/Peroni01/BalisageVol3-Peroni01.htmldoi:10.4242/BalisageVol3.Peroni01.

**Hilbert, M., Schonefeld, O. and Witt, A.** (2005). 'Making CONCUR work'. *Proceedings of Extreme Markup Languages 2005*. http://conferences.idealliance.org/extreme/html/2005/Witt01/EML2005Witt01.xml.

**Huitfeldt, C. and Sperberg-McQueen, C. M.** (2003). *TexMECS: An experimental markup meta-language for complex documents*. University of Bergen. http://decentius.aksis.uib.no/mlcd/2003/Papers/texmecs.html.

**Jagadish, H.V., Lakshmanan, L. V. S., Scannapieco, M., Srivastava, D. and Wiwatwattana, N.** (2004). 'Colorful XML: one hierarchy isn't enough'. *Proceedings of the 2004 ACM SIGMOD international conference on Management of data*. Paris (France), pp. 251-262. http://doi.acm.org/10.1145/1007568.1007598.

**Sperberg-McQueen, C.M. and Huitfeldt, C.** (1998). 'Concurrent Document Hierarchies in MECS and SGML'. *Literary and Linguistic Computing*. **14**: 29-42.

**Stührenberg, M. and Jettka, D.** (2009). 'A toolkit for multi-dimensional markup: The development of SGF to XStandoff'. *Proceedings of Balisage: The Markup Conference 2009*. Montréal (Canada), August 11-14, 2009. http://www.balisage.net/Proceedings/vol3/html/Stuhrenberg01/BalisageVol3-Stuhrenberg01.htmldoi:10.4242/BalisageVol3.Stuhrenberg01.

**Stührenberg, M. and Goecke, D.** (2008). 'SGF — An integrated model for multiple annotations and its application in a linguistic domain'. *Proceedings of Balisage: The Markup Conference 2008*. Montréal (Canada), August 12-15, 2008. http://www.balisage.net/Proceedings/vol1/html/Stuehrenberg01/BalisageVol1-Stuehrenberg01.htmldoi:10.4242/BalisageVol1.Stuehrenberg01.

**Sperberg-McQueen, C. M. and Burnard, L.** (1990). *Guidelines for the Encoding and Interchange of Machine-Readable Texts (TEI P1)*. Chicago, Oxford: Text Encoding Initiative.

**Witt, A., Lüngen, H., Sasaki, F. and Goecke, D.** (2005). 'Unification of XML Documents with Concurrent Markup'. *Literary and Linguistic Computing*. **20(1)**: 103-116.