Concurrent Markup for XML Documents

Patrick **Durusau**
Matthew **Brook O'Donnell**

Abstract

The implementation of concurrent markup by Durusau and O'Donnell (Extreme Markup 2001) relies upon related but separate principles. First, markup, commonly described in tree notation, is actually metadata about PCDATA. Second, the membership of any "atom of PCDATA" [1] in a given hierarchy can be recorded as metadata for that PCDATA. These two principles have allowed the authoring and querying of overlapping hierarchies using standard XML software.

This presentation moves beyond the use of text snippets to illustrate overlapping hierarchies and applies the authors' technique to one of the classics of Western literature, John Milton's *Paradise Lost*. This research has resulted in the first release of overlapping texts for experimentation on overlapping hierarchies and in a firmer theoretical foundation for current and future research on this topic.

Table of Contents

1. Introduction

This project grew out of the interest of the authors in biblical texts that have complex textual transmission histories, and for which overlapping views or hierarchies have developed over centuries of commentary and analysis. In addition modern analysis for linguistic, literary or textual study produces a complex of overlapping views. While a number of techniques exist for handling overlapping hierarchies, they all suffer from a variety of defects, ranging from non-implementation to inadequacy in querying of overlapping hierarchies. [2] The method developed by the authors allows the use of standard XML software for the editing, construction and querying of overlapping hierarchies.

The application of this method to a larger text (Paradise Lost) and the results of that experience, caused the authors' to examine the typology of overlapping hierarchies. What at first was seen as a technical issue of parsing and syntax, became apparent to have deeper significance for the encoding of texts.

[1] "Atom of PCDATA" is used in the sense of word tokens. That level of text division was chosen for the convenience of the researchers in constructing the sample data set. This is not a theoretical limit for this technique as it could be used to record the smallest addressable location in a text.

[2] See: http://www.sbl-site2.org/Extreme2001 for a fuller outline of those various defects.

2. Overlapping Hierarchies: A Partial Typology

Overlapping hierarchies (as traditionally understood) can be divided into thirteen separate cases.[4] Assuming the simplest case of two well formed and static elements, Element <a> and Element <b>, those may have the following structural relationships:

```
1. No-overlap

             <a>.....................</a>
<b>......</b>

2. Elements share one end/start point

             <a>.....................</a>
<b>.........</b>

3. 'Classic' overlap

            <a>.....................</a>
<b>.....................</b>

4. Elements share end point

            <a>.....................</a>
<b>.............................</b>

5. One element contained within the other

            <a>.....................</a>
<b>.........................................</b>

6. Elements share start point

            <a>.....................</a>
            <b>......</b>

7. Elements share both start and end points

            <a>.....................</a>
            <b>.....................</b>

8. Elements share start point

            <a>.....................</a>
            <b>.............................</b>

9. One element contained within the other

            <a>.....................</a>
                  <b>......</b>

10. Elements share end point

            <a>.....................</a>
                  <b>..............</b>

11. 'Classic' overlap
```

[4]The most complete analysis of overlapping hierarchies, including cases from overlapping pending edits (not treated here) can be found in David Durand's dissertation "Pamlimpest: Change-Oriented Concurrency Control for the Support of Collaborative Applications." [Durand 1999] The analysis presented here is based upon the authors' research and informal conversations with Steven DeRose and David Durand. Any errors or inaccuracies in the reporting or use of this analysis are solely the responsiblity of the authors.

```
            <a>.....................</a>
                  <b>.........................</b>

12.  Elements share one end/start point

            <a>.....................</a>
                                    <b>.......</b>

13. No overlap

           <a>.....................</a>
                                      <b>......</b>
```

Cases 1, 5, 9 and 13 are not diffcult for conventional markup, though the hierarchical nature of SGML/XML markup imposes a parent-child relationship in cases 5 and 9 that is not necessarily intended in the encoding view. Classic overlapping, shown by cases 3 and 11 are instances where one element does not properly "nest" inside a container element. Those cases were the initial focus of the research on this topic.

Classic overlapping as shown above (cases 3 and 11) has most commonly been encoded using milestones or stand-off markup. Carried to its ultimate conclusion, the former (milestones) could be used to encode an entire text within a single container element but with the loss of validated authoring. The latter (stand-off markup) can certainly create alternative tree structures, but only at the expense of creating a separate tree, which loses the information contained in the first one. (The lack of tools accessible to the average user is a practical concern for the use of stand-off markup. There are also difficulties associated with the query and retrieval of textual data across separate trees.)[5]

An alternative syntax has been developed by Michael Sperberg-McQueen and Claus Huitfeldt (MECS [6] and TexMECS[7]) to encode the traditional cases of overlapping hierarchies. The theoretical work of that project is quite extensive and recommended to anyone interested in issues of overlapping hierarchies. It should be noted that the TexMECS solution moves beyond standard SGML/XML syntax and therefore suffers from a lack of common implementations.

Successful treatment of classic overlapping brings a number of benefits to markup users and its importance should not be discounted. However, the hierarchical assumptions of classic overlapping color the solutions that will be sought and hence, ultimately influence the range of solutions that will be developed.

There are non-trivial cases of overlap that do not fall into the class of "classic" overlapping elements. Consider cases 2, 4, 7, 8, 10 and 12, where one or both opening and closing points for elements are shared in a text. The best case is number 7 since it highlights the question of what it means to overlap in a hierarchy.

```
7. Elements share both start and end points

            <a>.....................</a>
            <b>.....................</b>
```

This form of overlap illustrates that part of the problem of "overlapping" hierarchies lies in implied assumption of a single tree within which hierarchies conflict, overlap and nest. A cursory examination of most examples in this area, including the earlier work of the authors', reflects that assumption. While research is ongoing, it is tenatively suggested that further research on the question of whether structured markup requires a single tree or if viable alternatives can be devised within the SGML/XML framework is merited.

[5]See for example the treatment of such overlaps in the Text Encoding Initiative Guidelines (P3), at Chapter 31, Multiple Hierarchies [Sperberg-McQueen, 1994]

[6] **MECS - A Multi-Element Code System** by Claus Huitfeldt, http://helmer.hit.uib.no/claus/mecs/mecs.htm

[7] **TexMECS: An experimental markup meta-language for complex documents**, Claus Huitfeldt and C. M. Sperberg-McQueen, http://www.hit.uib.no/claus/mlcd/papers/texmecs.html

As a concrete example of the issues discussed in this section, consider two possible encodings of the first 23 words of the poem in Book 1 of *Paradise Lost*. The first marks line divisions as found in a number of common editions:

```
<line>Of Man's first disobedience, and the fruit</line>
<line>Of that forbidden tree whose mortal taste</line>
<line>Brought death into the World, and all our woe,</line>
```

The second analysis is of a more linguistic nature and marks sentence (`<s>`) and clause or clause section (marked with `<seg>`) divisions, indicated by the punctuation in the text.

```
<s>
 <seg>Of Man's first disobedience,</seg>
 <seg> and the fruit Of that forbidden tree whose mortal taste
Brought death into the World,</seg>
 <seg> and all our woe,</seg>
...
</s>
```

An examples of cases 6 and 8, where elements share a start point is the first `<line>` and `<seg>` elements:

```
<line><seg>Of Man's first disobedience,</seg> and the fruit</line>
```

To ensure well-formedness, conventional markup would encode the `<seg>` element within the `<line>`, which is in effect the same as cases 5 and 9 where one element is contained within another.

An example of classic overlap (cases 3 and 11 above) occurs with the second `<seg>` element:

```
<line>
 <seg>
  Of Man's first disobedience,
 </seg>
 <seg>
  and the fruit
</line>
<line>
  Of that forbidden tree whose mortal taste
</line>
<line>
  Brought death into the World,
 </seg>
```

No rearrangment of the nesting order of the `<line>` and `<seg>` elements will resolve these cases of overlap. Element segmentation is often suggested as a partial solution to the problem, but itself introduces processing difficulties. The final case of overlap in this 23 word example is the point after the last word "woe" where closing `<line>` and `<seg>` elements occur (cases 4 and 10 above).

3. BUVH: Basic Overview

As noted earlier, the method shown here constructs "bottom-up virtual hierarchies" (BUVH) [8] to encode the membership of any atom of #PCDATA in a particular hierarchy. A text is encoded in separate files, using different hierarchies, and then combined into a base file that uses XPath expressions to represent the membership of a par-

[8]The phrase "bottom-up virtual hierarchies" or BUVH, was coined by Michael Sperberg-McQueen in comments made during an earlier stage of this project.

ticular token of PCDATA in the respective hierarchies. The following example from the Milton's *Paradise Lost* illustrates the technique in operation.

In the traditional presentation of *Paradise Lost*, the text consists of "book," "div" and "line" divisions as follows (for the first 26 lines of the poem):

```
<div type="segment">
<line n="Bk.1.1">Of Man's first disobedience, and the fruit</line>
<line n="Bk.1.2">Of that forbidden tree whose mortal taste</line>
<line n="Bk.1.3">Brought death into the World, and all our woe,</line>
<line n="Bk.1.4">With loss of Eden, till one greater Man</line>
<line n="Bk.1.5">Restore us, and regain the blissful seat,</line>
<line n="Bk.1.6">Sing, Heavenly Muse, that, on the secret top</line>
<line n="Bk.1.7">Of Oreb, or of Sinai, didst inspire</line>
<line n="Bk.1.8">That Shepherd who first taught the chosen seed</line>
<line n="Bk.1.9">In the beginning how the heavens and earth</line>
<line n="Bk.1.10">Rose out of Chaos: or, if Sion hill</line>
<line n="Bk.1.11">Delight thee more, and Siloa's brook that flowed</line>
<line n="Bk.1.12">Fast by the oracle of God, I thence</line>
<line n="Bk.1.13">Invoke thy aid to my adventurous song,</line>
<line n="Bk.1.14">That with no middle flight intends to soar</line>
<line n="Bk.1.15">Above th' Aonian mount, while it pursues</line>
<line n="Bk.1.16">Things unattempted yet in prose or rhyme.</line>
<line n="Bk.1.17">And chiefly thou, O Spirit, that dost prefer</line>
<line n="Bk.1.18">Before all temples th' upright heart and pure,</line>
<line n="Bk.1.19">Instruct me, for Thou know'st; Thou from the first</line>
<line n="Bk.1.20">Wast present, and, with mighty wings outspread,</line>
<line n="Bk.1.21">Dove-like sat'st brooding on the vast Abyss,</line>
<line n="Bk.1.22">And mad'st it pregnant: what in me is dark</line>
<line n="Bk.1.23">Illumine, what is low raise and support;</line>
<line n="Bk.1.24">That, to the height of this great argument,</line>
<line n="Bk.1.25">I may assert Eternal Providence,</line>
<line n="Bk.1.26">And justify the ways of God to men.</line>
</div>
```

A simple content model for this hierarchy (which we refer to as the "textual" view) is:

```
<!ELEMENT book (div+)>
<!ELEMENT div (line+)>
<!ELEMENT line (#PCDATA)>
```

An alternative or additional analysis of the same portion of text is the "sentence" view, where divisions are made on the basis of clauses, sentences and other linguistic divisions. This analysis presents a different hierarchy altogether:

```
<div type="segment">
<s><seg>Of Man's first disobedience,</seg><seg> and the fruit
Of that forbidden tree whose mortal taste
Brought death into the World,</seg><seg> and all our woe,</seg><seg>
With loss of Eden,</seg><seg> till one greater Man
Restore us,</seg><seg> and regain the blissful seat,</seg><seg>
Sing,</seg><seg> Heavenly Muse,</seg><seg> that,</seg><seg> on the secret top
Of Oreb,</seg><seg> or of Sinai,</seg><seg> didst inspire
That Shepherd who first taught the chosen seed
In the beginning how the heavens and earth
Rose out of Chaos:</seg><seg> or,</seg><seg> if Sion hill
Delight thee more,</seg><seg> and Siloa's brook that flowed
Fast by the oracle of God,</seg><seg> I thence
Invoke thy aid to my adventurous song,</seg><seg>
```

```
That with no middle flight intends to soar
Above th' Aonian mount,</seg><seg> while it pursues
Things unattempted yet in prose or rhyme.</seg></s><s><seg>
And chiefly thou,</seg><seg> O Spirit,</seg><seg> that dost prefer
Before all temples th' upright heart and pure,</seg><seg>
Instruct me,</seg><seg> for Thou know'st;</seg><seg> Thou from the first
Wast present,</seg><seg> and,</seg><seg> with mighty wings outspread,
</seg><seg>
Dove-like sat'st brooding on the vast Abyss,</seg><seg>
And mad'st it pregnant:</seg><seg> what in me is dark
Illumine,</seg><seg> what is low raise and support;</seg><seg>
That,</seg><seg> to the height of this great argument,</seg><seg>
I may assert Eternal Providence,</seg><seg>
And justify the ways of God to men.</seg></s></div>
```

The content model for the "sentence" hierarchy is:

```
<!ELEMENT div (s+)>
<!ELEMENT s (seg+)>
<!ELEMENT seg (#PCDATA)>
```

Note that both the textual and sentence views use a `div` element with a `type` attribute with the value "segment". It might be more precise to label the `div` element belonging to the sentence hierarchy as a "paragraph", although this would be more interpretive. However, the two `div` elements represent an instance of case 7 where they coincide.

A third view or hierarchy is the "page" view that followsing the physical page divisions in a standard printed edition [Raffel, 1999] of the poem (the presentational alternative of traditional concur in SGML):

```
<div type="book" n="1">
<page n="135">
<line n="Bk.1.1">Of Man's first disobedience, and the fruit</line>
<line n="Bk.1.2">Of that forbidden tree whose mortal taste</line>
<line n="Bk.1.3">Brought death into the World, and all our woe,</line>
<line n="Bk.1.4">With loss of Eden, till one greater Man</line>
<line n="Bk.1.5">Restore us, and regain the blissful seat,</line>
<line n="Bk.1.6">Sing, Heavenly Muse, that, on the secret top</line>
<line n="Bk.1.7">Of Oreb, or of Sinai, didst inspire</line>
<line n="Bk.1.8">That Shepherd who first taught the chosen seed</line>
<line n="Bk.1.9">In the beginning how the heavens and earth</line>
<line n="Bk.1.10">Rose out of Chaos: or, if Sion hill</line>
<line n="Bk.1.11">Delight thee more, and Siloa's brook that flowed</line>
<line n="Bk.1.12">Fast by the oracle of God, I thence</line>
<line n="Bk.1.13">Invoke thy aid to my adventurous song,</line>
<line n="Bk.1.14">That with no middle flight intends to soar</line>
<line n="Bk.1.15">Above th' Aonian mount, while it pursues</line>
<line n="Bk.1.16">Things unattempted yet in prose or rhyme.</line>
<line n="Bk.1.17">And chiefly thou, O Spirit, that dost prefer</line>
<line n="Bk.1.18">Before all temples th' upright heart and pure,</line>
<line n="Bk.1.19">Instruct me, for Thou know'st; Thou from the first</line>
<line n="Bk.1.20">Wast present, and, with mighty wings outspread,</line>
<line n="Bk.1.21">Dove-like sat'st brooding on the vast Abyss,</line>
<line n="Bk.1.22">And mad'st it pregnant: what in me is dark</line>
<line n="Bk.1.23">Illumine, what is low raise and support;</line>
<line n="Bk.1.24">That, to the height of this great argument,</line>
<line n="Bk.1.25">I may assert Eternal Providence,</line>
<line n="Bk.1.26">And justify the ways of God to men.</line>
<line n="Bk.1.27">Say first--for Heaven hides nothing from thy</line>
<line n="Bk.1.28"> view,</line>
</page>
```

The content model for the "page" hierarchy is:

```
<!ELEMENT div (page+)>
<!ELEMENT page (line+)>
<!ELEMENT line (#PCDATA)>
```

Although both the textual and page views include a `line` element and they often coincide (case 7) they are separate logical entities. In addition there are variations, for instance line "Bk.1.28" in the page view consists of a single word.

All of these are legitimate "views" (or "hierarchies" as the term is commonly used) of the text. But the problem arises of how to express all three in such a way to allow for querying across the hierarchies and to avoid ill-formed XML.

The use of XPath expressions to record the hierarchy information for each PCDATA token as attributes allows the preservation of the membership information in a particular hierarchy, while allowing the use of traditional (and widely available) XML tools. The base file constructed for the hierarchies listed above reads for the first 23 words of the poem as follows:

```
<baseFile xmlns:ln="urn:line-divs-1674"
          xmlns:sn="urn:sentence-segs-1674"
          xmlns:pg="urn:pages-lines-1674"
>
   <w id="w296"
      ln:text="/text/div[1][@type='book'][@n='1']
        /div[2][@type='segment']/line[1][@n='Bk.1.1']/*[1]"
      sn:text="/text/div[1][@type='book'][@n='1']
        /div[2][@type='segment']/s[1]/seg[1]/*[1]"
      pg:text="/text/div[1][@type='book'][@n='1']
        /div[2][@type='segment']/page[1][@n='135']/
        line[1][@n='Bk.1.1']/*[1]"
   >Of</w>
   <w id="w297"
      ln:text="/text/div[1][@type='book'][@n='1']
        /div[2][@type='segment']/line[1][@n='Bk.1.1']/*[2]"
      sn:text="/text/div[1][@type='book'][@n='1']
        /div[2][@type='segment']/s[1]/seg[1]/*[2]"
      pg:text="/text/div[1][@type='book'][@n='1']
        /div[2][@type='segment']/page[1][@n='135']/
        line[1][@n='Bk.1.1']/*[2]"
   >Man's</w>
   <w id="w298"
      ln:text="/text/div[1][@type='book'][@n='1']
        /div[2][@type='segment']/line[1][@n='Bk.1.1']/*[3]"
      sn:text="/text/div[1][@type='book'][@n='1']
        /div[2][@type='segment']/s[1]/seg[1]/*[3]"
      pg:text="/text/div[1][@type='book'][@n='1']
        /div[2][@type='segment']/page[1][@n='135']/
        line[1][@n='Bk.1.1']/*[3]"
   >first</w>
   <w id="w299"
      ln:text="/text/div[1][@type='book'][@n='1']
        /div[2][@type='segment']/line[1][@n='Bk.1.1']/*[4]"
      sn:text="/text/div[1][@type='book'][@n='1']
        /div[2][@type='segment']/s[1]/seg[1]/*[4]"
      pg:text="/text/div[1][@type='book'][@n='1']
        /div[2][@type='segment']/page[1][@n='135']/
        line[1][@n='Bk.1.1']/*[4]"
   >disobedience,</w>
   <w id="w300"
      ln:text="/text/div[1][@type='book'][@n='1']
```

```
      /div[2][@type='segment']/line[1][@n='Bk.1.1']/*[5]"
   sn:text="/text/div[1][@type='book'][@n='1']
     /div[2][@type='segment']/s[1]/seg[2]/*[1]"
   pg:text="/text/div[1][@type='book'][@n='1']
     /div[2][@type='segment']/page[1][@n='135']/
     line[1][@n='Bk.1.1']/*[5]"
>and</w>
<w id="w301"
   ln:text="/text/div[1][@type='book'][@n='1']
     /div[2][@type='segment']/line[1][@n='Bk.1.1']/*[6]"
   sn:text="/text/div[1][@type='book'][@n='1']
     /div[2][@type='segment']/s[1]/seg[2]/*[2]"
   pg:text="/text/div[1][@type='book'][@n='1']
     /div[2][@type='segment']/page[1][@n='135']/
     line[1][@n='Bk.1.1']/*[6]"
>the</w>
<w id="w302"
   ln:text="/text/div[1][@type='book'][@n='1']
     /div[2][@type='segment']/line[1][@n='Bk.1.1']/*[7]"
   sn:text="/text/div[1][@type='book'][@n='1']
     /div[2][@type='segment']/s[1]/seg[2]/*[3]"
   pg:text="/text/div[1][@type='book'][@n='1']
     /div[2][@type='segment']/page[1][@n='135']/
     line[1][@n='Bk.1.1']/*[7]"
>fruit</w>
<w id="w303"
   ln:text="/text/div[1][@type='book'][@n='1']
     /div[2][@type='segment']/line[2][@n='Bk.1.2']/*[1]"
   sn:text="/text/div[1][@type='book'][@n='1']
     /div[2][@type='segment']/s[1]/seg[2]/*[4]"
   pg:text="/text/div[1][@type='book'][@n='1']
     /div[2][@type='segment']/page[1][@n='135']
     /line[2][@n='Bk.1.2']/*[1]"
>Of</w>
<w id="w304"
   ln:text="/text/div[1][@type='book'][@n='1']
     /div[2][@type='segment']/line[2][@n='Bk.1.2']/*[2]"
   sn:text="/text/div[1][@type='book'][@n='1']
     /div[2][@type='segment']/s[1]/seg[2]/*[5]"
   pg:text="/text/div[1][@type='book'][@n='1']
     /div[2][@type='segment']/page[1][@n='135']
     /line[2][@n='Bk.1.2']/*[2]"
>that</w>
<w id="w305"
   ln:text="/text/div[1][@type='book'][@n='1']
     /div[2][@type='segment']/line[2][@n='Bk.1.2']/*[3]"
   sn:text="/text/div[1][@type='book'][@n='1']
     /div[2][@type='segment']/s[1]/seg[2]/*[6]"
   pg:text="/text/div[1][@type='book'][@n='1']
     /div[2][@type='segment']/page[1][@n='135']/
     line[2][@n='Bk.1.2']/*[3]"
>forbidden</w>
<w id="w306"
   ln:text="/text/div[1][@type='book'][@n='1']
     /div[2][@type='segment']/line[2][@n='Bk.1.2']/*[4]"
   sn:text="/text/div[1][@type='book'][@n='1']
     /div[2][@type='segment']/s[1]/seg[2]/*[7]"
   pg:text="/text/div[1][@type='book'][@n='1']
     /div[2][@type='segment']/page[1][@n='135']/
     line[2][@n='Bk.1.2']/*[4]"
>tree</w>
<w id="w307"
   ln:text="/text/div[1][@type='book'][@n='1']
```

```
      /div[2][@type='segment']/line[2][@n='Bk.1.2']/*[5]"
   sn:text="/text/div[1][@type='book'][@n='1']
     /div[2][@type='segment']/s[1]/seg[2]/*[8]"
   pg:text="/text/div[1][@type='book'][@n='1']
     /div[2][@type='segment']/page[1][@n='135']/
     line[2][@n='Bk.1.2']/*[5]"
>whose</w>
<w id="w308"
   ln:text="/text/div[1][@type='book'][@n='1']
     /div[2][@type='segment']/line[2][@n='Bk.1.2']/*[6]"
   sn:text="/text/div[1][@type='book'][@n='1']
     /div[2][@type='segment']/s[1]/seg[2]/*[9]"
   pg:text="/text/div[1][@type='book'][@n='1']
     /div[2][@type='segment']/page[1][@n='135']/
     line[2][@n='Bk.1.2']/*[6]"
>mortal</w>
<w id="w309"
   ln:text="/text/div[1][@type='book'][@n='1']
     /div[2][@type='segment']/line[2][@n='Bk.1.2']/*[7]"
   sn:text="/text/div[1][@type='book'][@n='1']
     /div[2][@type='segment']/s[1]/seg[2]/*[10]"
   pg:text="/text/div[1][@type='book'][@n='1']
     /div[2][@type='segment']/page[1][@n='135']/
     line[2][@n='Bk.1.2']/*[7]"
>taste</w>
<w id="w310"
   ln:text="/text/div[1][@type='book'][@n='1']
     /div[2][@type='segment']/line[3][@n='Bk.1.3']/*[1]"
   sn:text="/text/div[1][@type='book'][@n='1']
     /div[2][@type='segment']/s[1]/seg[2]/*[11]"
   pg:text="/text/div[1][@type='book'][@n='1']
     /div[2][@type='segment']/page[1][@n='135']/
     line[3][@n='Bk.1.3']/*[1]"
>Brought</w>
<w id="w311"
   ln:text="/text/div[1][@type='book'][@n='1']
     /div[2][@type='segment']/line[3][@n='Bk.1.3']/*[2]"
   sn:text="/text/div[1][@type='book'][@n='1']
     /div[2][@type='segment']/s[1]/seg[2]/*[12]"
   pg:text="/text/div[1][@type='book'][@n='1']
     /div[2][@type='segment']/page[1][@n='135']/
     line[3][@n='Bk.1.3']/*[2]"
>death</w>
<w id="w312"
   ln:text="/text/div[1][@type='book'][@n='1']
     /div[2][@type='segment']/line[3][@n='Bk.1.3']/*[3]"
   sn:text="/text/div[1][@type='book'][@n='1']
     /div[2][@type='segment']/s[1]/seg[2]/*[13]"
   pg:text="/text/div[1][@type='book'][@n='1']
     /div[2][@type='segment']/page[1][@n='135']/
     line[3][@n='Bk.1.3']/*[3]"
>into</w>
<w id="w313"
   ln:text="/text/div[1][@type='book'][@n='1']
     /div[2][@type='segment']/line[3][@n='Bk.1.3']/*[4]"
   sn:text="/text/div[1][@type='book'][@n='1']
     /div[2][@type='segment']/s[1]/seg[2]/*[14]"
   pg:text="/text/div[1][@type='book'][@n='1']
     /div[2][@type='segment']/page[1][@n='135']/
     line[3][@n='Bk.1.3']/*[4]"
>the</w>
<w id="w314"
   ln:text="/text/div[1][@type='book'][@n='1']
```

```
        /div[2][@type='segment']/line[3][@n='Bk.1.3']/*[5]"
      sn:text="/text/div[1][@type='book'][@n='1']
        /div[2][@type='segment']/s[1]/seg[2]/*[15]"
      pg:text="/text/div[1][@type='book'][@n='1']
        /div[2][@type='segment']/page[1][@n='135']/
        line[3][@n='Bk.1.3']/*[5]"
   >World,</w>
   <w id="w315"
      ln:text="/text/div[1][@type='book'][@n='1']
        /div[2][@type='segment']/line[3][@n='Bk.1.3']/*[6]"
      sn:text="/text/div[1][@type='book'][@n='1']
        /div[2][@type='segment']/s[1]/seg[3]/*[1]"
      pg:text="/text/div[1][@type='book'][@n='1']
        /div[2][@type='segment']/page[1][@n='135']/
        line[3][@n='Bk.1.3']/*[6]"
   >and</w>
   <w id="w316"
      ln:text="/text/div[1][@type='book'][@n='1']
        /div[2][@type='segment']/line[3][@n='Bk.1.3']/*[7]"
      sn:text="/text/div[1][@type='book'][@n='1']
        /div[2][@type='segment']/s[1]/seg[3]/*[2]"
      pg:text="/text/div[1][@type='book'][@n='1']
        /div[2][@type='segment']/page[1][@n='135']/
        line[3][@n='Bk.1.3']/*[7]"
   >all</w>
   <w id="w317"
      ln:text="/text/div[1][@type='book'][@n='1']
        /div[2][@type='segment']/line[3][@n='Bk.1.3']/*[8]"
      sn:text="/text/div[1][@type='book'][@n='1']
        /div[2][@type='segment']/s[1]/seg[3]/*[3]"
      pg:text="/text/div[1][@type='book'][@n='1']
        /div[2][@type='segment']/page[1][@n='135']/
        line[3][@n='Bk.1.3']/*[8]"
   >our</w>
   <w id="w318"
      ln:text="/text/div[1][@type='book'][@n='1']
        /div[2][@type='segment']/line[3][@n='Bk.1.3']/*[9]"
      sn:text="/text/div[1][@type='book'][@n='1']
        /div[2][@type='segment']/s[1]/seg[3]/*[4]"
      pg:text="/text/div[1][@type='book'][@n='1']
        /div[2][@type='segment']/page[1][@n='135']/
        line[3][@n='Bk.1.3']/*[9]"
   >woe,</w>
</baseFile>
```

Each of the three hierarchies is placed in its own namespace with the prefixes: `ln` for the textual view, `sn` for the sentence view and `pg` for the page view. Each #PCDATA atom in the text is enclosed in a `w` element and given a identifier (the `id` attribute). The XPath expression for the #PCDATA atom in each hierarchy is then encoded in an namespace qualified attribute.

To demonstrate the advantages of this approach, consider the example of overlap cases 6 and 8 discussed above:

```
<line><seg>Of Man's first disobedience,</seg> and the fruit</line>
```

here the two elements share a common start point. Conventional in-line markup forces the encoder to develop a nesting order, whereas the BUVH version keeps the hierarchies physically (in terms of markup) separate:

```
<w id="w296"
   ln:text="/text/div[1][@type='book'][@n='1']
     /div[2][@type='segment']/line[1][@n='Bk.1.1']/*[1]"
```

```
   sn:text="/text/div[1][@type='book'][@n='1']
     /div[2][@type='segment']/s[1]/seg[1]/*[1]"
>Of</w>
```

An examination of the other points and types of overlap in the first 23 words of Book 1, which were discussed above, in the BUVH version reveals the power of this encoding.

In practical terms it should be noted that the three files representing chapter one of Paradise Lost totaled 164,793 bytes, while the resulting base file was 1,721,382, some ten times larger. On a 1.2 Ghz laptop with 1 GB of RAM, Saxon took a little over 7 minutes to generate the base file included in the test file distribution. The redundancy of hierarchy information (and the verboseness of XPath expressions) in part contributes to the rapid growth in file size.

The recording of such detailed information allows the exploration of issues such as depth of commonality in hierarchies, the nature of overlaps between hierarchies and other issues touching upon the nature of overlapping hierarchies. The strongest feature of this technique remains its reliance on standard XML tools for experimentation with and exploration of overlapping hierarchy issues.

4. Overlapping Hierarchies in Paradise Lost

Paradise Lost, like all literary works, is a very complex text that has been the object of scholarly study almost since it original publication. With apologies to any Milton scholars, the authors have encoded only the most basic overlapping hierarchies (in the sense of structural and not conceptual overlap) for this project. It is hoped that further research on this particular text will enlist the aid of Milton scholars to identify and encode conceptual overlap. Such encodings would be useful in testing any theoretical models that include conceptual overlap.

Paradise Lost was originally published in ten "books" in 1667. It was substantially revised, with several books being split into separate books and the addition of introductory materials at the beginning of each book, denoted as "arguments" and republished in 1674. The later edition is the one most often reprinted in whole or in part in various collections of Milton's writings. In terms of genre, *Paradise Lost* is an epic poem, which has a complex interplay between various parts of the text.

The overlapping hierarchies used herein are based upon the 1674 edition of the text and represent three of the most common overlapping hierarchies: line breaks in the text versus breaks in the text signalled with periods, semi-colons, colons and simple indicators of clause divisions. Since the presentation "view" was one of the original uses for the concur feature of SGML, an alternative hierarchy based upon a modern edition's page and line breaks was also constructed. Beyond validation of the files themselves, it is not warranted that the markup would be useful for actual Milton studies.

Electronic versions of the files prepared for both versions of the Milton text are available for download to facilitate the evaluation of this technique as well as comparison of other strategies for encoding overlapping hierarchies. These materials are being released subject only to the condition that the origin of the materials be acknowledged in any subsequent use or publication. The XSLT scripts and our current results are also available for downloading.[9]

One interesting side note is that in the encoding process, one tends to read the text more carefully than as an undergraduate and discovers Adams' debate with himself over whether to eat the fruit that has been offered to him by Eve. Contrasted with the lie later told in the Garden, Adam knows the origin of the fruit and makes a concious decision to join Eve in sin.

5. Observations and Further Investigations

The application of the BUVH to *Paradise Lost* has led to several observations about both this technique in particular and raised questions about overlapping markup in general. First, while perhaps computationally intensive, the verbosity of the current approach permits the evaluation of the relationships between hierarchies both visually as

[9] http://www.sbl-site2.org/XMLEurope2002

well as computationally. One of the open questions for further investigation is the use of this technique with other example texts that are said to exhibit overlapping markup. Since BUVH allows for easy authoring of such overlapping hierarchies with standard editors, it is hoped that more example texts will appear for investigation of cases of overlap. Rather than interesting but small examples, investigators will be able to test theories of overlapping hierarchies against significant collections of materials.

Second, the texts (and scripts) released with this paper provide a set of test cases for the investigation of overlapping hierarchies in texts. Investigators are urged to prepare other substantial sample texts to provide a wide range of material for use in such investigations.

Finally, the investigation into a typology of overlapping hierarchies, inspired in part by the experimental results from this research, lead to the suggestion that the focus on single trees in the investigation of overlapping hierarchies may be misplaced. It may be the case that the focus on single trees in encoding technology has led to a poorer view of texts. That is not to deny that single trees have been very useful and continue to be so. However, we wish to raise the issue that proceeding from texts to encoding may lead to different encoding techniques than from single trees to encoding texts. The technique presented provides one opportunity to investigate such issues while using traditional single tree based techniques.[10]

Bibliography

[Barnard, 1988] Barnard, David and Cherly A. Fraser, George M. Logan **Generalized Markup for Literary Texts**. *Literary and Linguistic Computing* Vol. 3, No. 1, 1988, 26-31

[Barnard, 1988a] Barnard, David and Ron Hayter, Maria Karababa, George Logan, John McFadden. **SGML-Based Markup Literary Texts: Two Problems and Some Solutions**. *Computers and the Humanities* 22 (1988) 265-276

[Barnard, 1992] Barnard, David, Lou Burnard, Jean-Pierre Gaspart Lynne Price C. M. Sperberg-McQueen Nino Varile **Notes on SGML Solutions To Markup Problems** TEI Metalanguage Working Group paper, TEI MLW18, dated April 16, 1992

[Caton, 2001] Caton, Paul, **Markup's Current Imbalance**, Markup Languages: Theory and Practices, volume 3, number 1, Winter 2001

[Coombs, 1987] Coombs, J. and Allen Renear, Steve DeRose. **Markup Systems and the Future of Scholarly Text Processing** http://www.oasis-open.org/cover/coombs.html

[Durand 1999] Durand, David G., 1999. **Palimpest: Change-Oriented Concurrency Control for the Support of Collaborative Applications**, Dissertation, Boston University. (available online at: http://cs-people.bu.edu/dgd

[Driscoll, 2001] Driscoll, M. J. **Encoding Old Norse-Icelandic primary sources using TEI-conformant SGML/XML: A handbook** http://www.hum.ku.dk/ami/handbook/

[Huitfeldt, 1998] Huitfeldt, Claus **MECS - A Multi-Element Code System** http://helmer.hit.uib.no/claus/mecs/mecs.htm

[Huitfeldt, 2001] Huitfeldt, Claus and C. M. Sperberg-McQueen. **TexMECS: An experimental markup meta-language for complex documents** http://www.hit.uib.no/claus/mlcd/papers/texmecs.html

[Huitfeldt, 1995] Huitfeldt, Claus, **Multi-Dimensional Texts in a One-Dimensional Medium** http://www.hit.uib.no/wab/el_texts/no5/huitfeldt.htm

[Mah, 1998] Mah, Carol E. **Introduction to Encoding: A Tutorial For New Encoders** http://www.wwp.brown.edu/encoding/training/intro/intro.html

[10] Paul Caton's recent article **Markup's Current Imbalance** is a non-technical introduction to the issues surrounding overlapping markup and the impact of the single tree view on encoding practices.[Caton, 2001]

[Raffel, 1999] Raffel, Burton (ed.), **The Annotated Milton**, Bantam Books, New York, London, 1999.

[Renear, 1995] Renear, Allen and Elli Mylonas, David Durand. **Refining our Notion of What Text Really Is: The Problem of Overlapping Hierarchies** http://www.stg.brown.edu/resources/stg/monographs/ohco.html

[Sperberg-McQueen, 1991] Sperberg-McQueen, C.M. **Text in the Electronic Age: Textual Study and Text Encoding, with Examples from Medieval Texts** *Literary and Linguistic Computing* Vol. 6, No. 1, 1991, 34-46

[Sperberg-McQueen, 1994] Sperberg-McQueen, C.M. and Lou Burnard (eds) **Guidelines for Text Encoding and Interchange** (P3) Chicago and Oxford: Text Encoding Initiative

[Sperberg-McQueen, 1999] Sperberg-McQueen, C.M. and Claus Huitfeldt **GODDAG: A Data Structure for Overlapping Hierarchies** http://www.uic.edu/~cmsmcq/tech/goddag.html

Glossary

BUVH	Bottom-up Virtual Hierarchies

Biography

Wednesday, 22 May 16.00

Patrick **Durusau**
Society of Biblical Literature
USA

Patrick Durusau is the Director of Research and Development for the Society of Biblical Literature (SBL). His primary research interests include the use of XML for the encoding and analysis of biblical and Ancient Near Eastern texts.

Matthew **Brook O'Donnell**
OpenText.org
USA

Matthew Brook O'Donnell is Director of Research and Development for OpenText.org, an initiative to develop XML-based tools and resources for linguistic analysis. His research interests include corpus linguistics, text encoding and the linguistics analysis of ancient Greek.