Visualizing Overlapping Hierarchies in Textual Markup

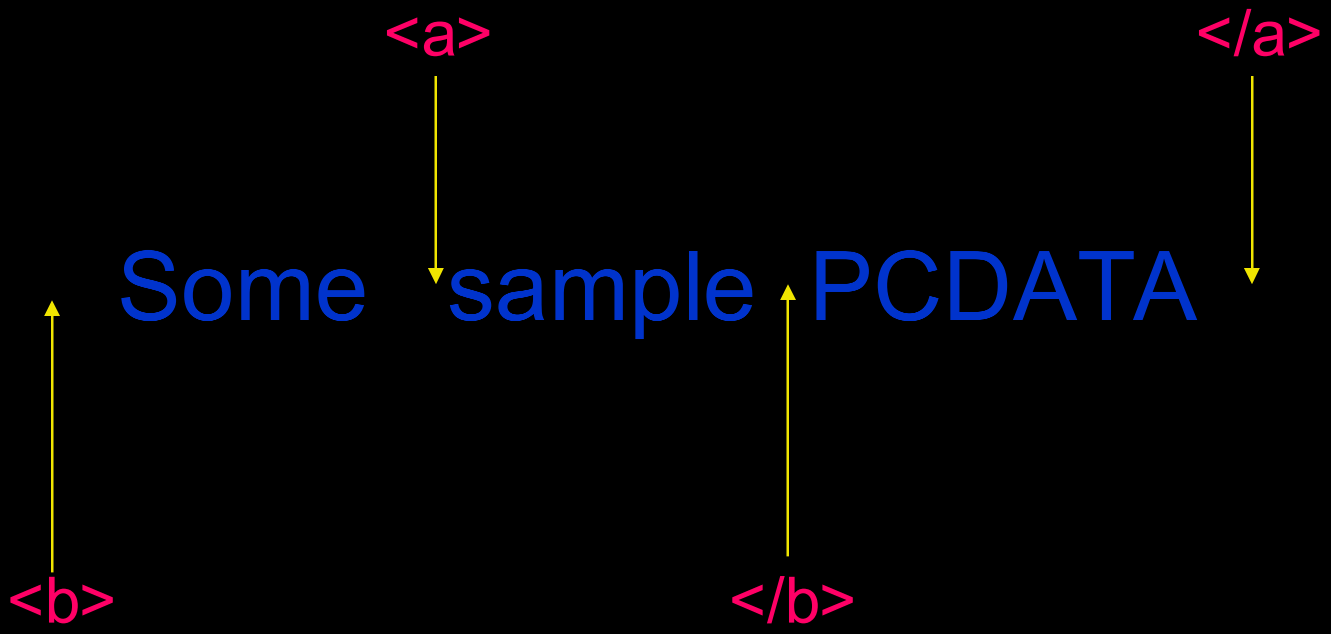
Patrick Durusau (*Society of Biblical Literature*)
pdurusau@emory.edu

Matthew Brook O’Donnell (*OpenText.org*)
matt@opentext.org

Background: Overlapping Hierarchies

- Features that do not nest within other features (TEI)
  - Speech beginning in a paragraph and spanning others
  - Verse when tagging formal structure and syntactic structure
  - Physical versus logical structure of a text
  - Versioning (not mentioned in TEI)

Weakness of Trees 1

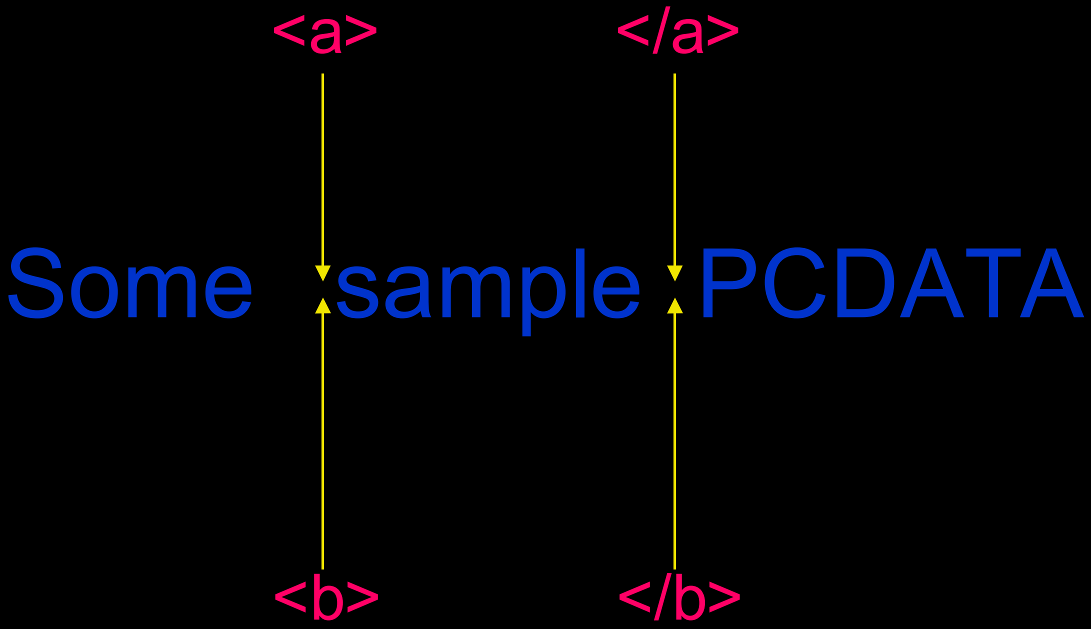


<a>
</a>

Some sample PCDATA

<b>
</b>

Weakness of Trees 2



<a> </a>

Some sample PCDATA

<b> </b>

Prior Solutions

- Concur (not implemented)
- Fragmentation/Virtual joins
- Milestones
- Multiple versions
- Non-SGML/XML markup
- Stand-off markup

BUVH: Bottom Up Virtual Hierarchies

- Two Principles:
  - Markup is metadata about PCDATA
  - Atoms of PCDATA have “membership” in markup constructs
- Uses XPath to record the membership information for various hierarchies

BUVH: Example

```
<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>
```

BUVH: Advantages

- Any structure
- Querying across structures
- XML software
- XML syntax

BUVH: Disadvantages

- Effective querying requires detailed knowledge of the hierarchies
- Size/Processing Requirements (10X increase in file size with 3 hierarchies)
- Visual assessment difficult even for markup specialists

Visualizing with BUVH

- BUVH has recorded all nodes and their PCDATA members
- Recorded information used to construct a directed acyclic graph
- XPath information converted into the *DOT* graph language (further information on *Dot*, the graph program, see:)

BUVH and Dot

- First pass: Simple parsing into nodes
- Second pass: Consolidate common nodes (note appearance of localized overlap)
- Third pass: Refine appearance of nodes
- Forth pass: Refine rankings (if necessary)

Visualizing BUVHs

- Assign shapes to nodes
- Collapse common nodes
- Select fragments (due to visual complexity)
- Use labels to represent text

Future Work

- Automatic generation of VRML with Dot
- Collapsing of nodes in BUVH
- Exploring the nature of overlapping markup
- Linking of text to VRML