TCS tcs: Tata Consultancy Services trash compactor script

Charlie Halpern-Hamu
Senior Solutions Architect,
Tata Consultancy Services
charlie.hamu@tcs.com

International Symposium on
Quality Assurance and Quality Control in XML,
Monday, 2012 August 6, Hotel Europa, Montreal

What are we talking about today?

- A program that
  creates sample documents
  by sampling documents

- Garbage In, Garbage Out
- So build a trash compactor script

- Compress a coal mine into a diamond

Reasons for creating samples

- Document analysis
- Vocabulary redesign
- Conversion design
- Storage design
- Editorial system design
- Transformation

- Coordination between parties
  = agreed interface

Approaches to creating samples

- Craft
  - Artful, thoughtful, manual
  - Painfully labor-intensive
- Mad Libs
  - Random and disconnected from reality
  - Hard for humans to comprehend and use
- Curation
  - Thoughtful choice of real data from larger set
  - XML is like war: long stretches of boredom punctuated by moments of awkwardness

Goals

- “Complete” results
- Plausible results
- Efficient on large corpora

- Good defaults
- Single dial

- Ease over perfection
- Use, or perhaps better, intuit schema

Algorithm

- Annotate with signatures

```
<em tcs:signature="para em">cat</em>
<em tcs:signature="para em">dog</em>
```

- Mark unique elements

```
<em tcs:mark="keep">cat</em>
<em>dog</em>
```

- Mark wrapping elements
- Mark required children, desired inlines
- Prune unneeded elements

```
<em>cat</em>
```

Example

```
<doc><title>Document Title</title>
    <p>First <em rend='bold'>paragraph</em>.</p>
    <section><title>Section One</title>
        <p>Second paragraph.</p>
        <p>Third paragraph.</p></section>
    <section><title>Section Two</title>
        <p>Fourth paragraph.</p></section>
    <section><title>Section Three</title>
        <p>Fifth <em>paragraph</em>.</p>
        <p>Sixth <em rend='ital'>para</em>.</p>
        <p>Seventh paragraph.</p></section></doc>
```

Example

```
<doc><title>Document Title</title>
    <p>First <em rend='bold'>paragraph</em>.</p>
    <section><title>Section One</title>
        <p>Second paragraph.</p>
        <p>Third paragraph.</p></section>
    <section><title>Section Two</title>
        <p>Fourth paragraph.</p></section>
    <section><title>Section Three</title>
        <p>Fifth <em>paragraph</em>.</p>
        <p>Sixth <em rend='ital'>para</em>.</p>
        <p>Seventh paragraph.</p></section></doc>
```

Example

```
<doc><title>Document Title</title>
    <p>First <em rend='bold'>paragraph</em>.</p>
    <section><title>Section One</title>
        <p>Second paragraph.</p>
        <p>Third paragraph.</p></section>
    <section><title>Section Two</title>
        <p>Fourth paragraph.</p></section>
    <section><title>Section Three</title>
        <p>Fifth <em>paragraph</em>.</p>
        <p>Sixth <em rend='ital'>para</em>.</p>
        <p>Seventh paragraph.</p></section></doc>
```

Example

```
<doc><title>Document Title</title>
    <p>First <em rend='bold'>paragraph</em>.</p>
    <section><title>Section One</title>
        <p>Second paragraph.</p>
        <p>Third paragraph.</p></section>
    <section><title>Section Two</title>
        <p>Fourth paragraph.</p></section>
    <section><title>Section Three</title>
        <p>Fifth <em>paragraph</em>.</p>
        <p>Sixth <em rend='ital'>para</em>.</p>
        <p>Seventh paragraph.</p></section></doc>
```

Example

```
<doc><title>Document Title</title>
    <p>First <em rend='bold'>paragraph</em>.</p>
    <section><title>Section One</title>
        <p>Second paragraph.</p>
        <p>Third paragraph.</p></section>
    <section><title>Section Two</title>
        <p>Fourth paragraph.</p></section>
    <section><title>Section Three</title>
        <p>Fifth <em>paragraph</em>.</p>
        <p>Sixth <em rend='ital'>para</em>.</p>
        <p>Seventh paragraph.</p></section></doc>
```

Example

```
<doc><title>Document Title</title>
    <p>First <em rend='bold'>paragraph</em>.</p>
    <section><title>Section One</title>
        <p>Second paragraph.</p>
        <p>Third paragraph.</p></section>
    <section><title>Section Two</title>
        <p>Fourth paragraph.</p></section>
    <section><title>Section Three</title>
        <p>Fifth <em>paragraph</em>.</p>
        <p>Sixth <em rend='ital'>para</em>.</p>
        <p>Seventh paragraph.</p></section></doc>
```

Example

```
<doc><title>Document Title</title>
    <p>First <em rend='bold'>paragraph</em>.</p>
    <section><title>Section One</title>
        <p>Second paragraph.</p>
        <p>Third paragraph.</p></section>
    <section><title>Section Two</title>
        <p>Fourth paragraph.</p></section>
    <section><title>Section Three</title>
        <p>Fifth <em>paragraph</em>.</p>
        <p>Sixth <em rend='ital'>para</em>.</p>
        <p>Seventh paragraph.</p></section></doc>
```

Example

```
<doc><title>Document Title</title>
    <p>First <em rend='bold'>paragraph</em>.</p>
    <section><title>Section One</title>
        <p>Second paragraph.</p>
        <p>Third paragraph.</p></section>
    <section><title>Section Two</title>
        <p>Fourth paragraph.</p></section>
    <section><title>Section Three</title>
        <p>Fifth <em>paragraph</em>.</p>
        <p>Sixth <em rend='ital'>para</em>.</p>
        <p>Seventh paragraph.</p></section></doc>
```

Example

```
<doc><title>Document Title</title>
    <p>First <em rend='bold'>paragraph</em>.</p>
    <section><title>Section One</title>
        <p>Second paragraph.</p>
        <p>Third paragraph.</p></section>
    <section><title>Section Two</title>
        <p>Fourth paragraph.</p></section>
    <section><title>Section Three</title>
        <p>Fifth <em>paragraph</em>.</p>
        <p>Sixth <em rend='ital'>para</em>.</p>
        <p>Seventh paragraph.</p></section></doc>
```

Example

```
<doc><title>Document Title</title>
    <p>First <em rend='bold'>paragraph</em>.</p>
    <section><title>Section One</title>
        <p>Second paragraph.</p>
        <p>Third paragraph.</p></section>
    <section><title>Section Two</title>
        <p>Fourth paragraph.</p></section>
    <section><title>Section Three</title>
        <p>Fifth <em>paragraph</em>.</p>
        <p>Sixth <em rend='ital'>para</em>.</p>
        <p>Seventh paragraph.</p></section></doc>
```

Example

```
<doc><title>Document Title</title>
    <p>First <em rend='bold'>paragraph</em>.</p>
    <section><title>Section One</title>
        <p>Second paragraph.</p>
        <p>Third paragraph.</p></section>
    <section><title>Section Two</title>
        <p>Fourth paragraph.</p></section>
    <section><title>Section Three</title>
        <p>Fifth <em>paragraph</em>.</p>
        <p>Sixth <em rend='ital'>para</em>.</p>
        <p>Seventh paragraph.</p></section></doc>
```

Example

```
<doc><title>Document Title</title>
    <p>First <em rend='bold'>paragraph</em>.</p>
    <section><title>Section One</title>
        <p>Second paragraph.</p>
        <p>Third paragraph.</p></section>
    <section><title>Section Two</title>
        <p>Fourth paragraph.</p></section>
    <section><title>Section Three</title>
        <p>Fifth <em>paragraph</em>.</p>
        <p>Sixth <em rend='ital'>para</em>.</p>
        <p>Seventh paragraph.</p></section></doc>
```

Example

```
<doc><title>Document Title</title>
    <p>First <em rend='bold'>paragraph</em>.</p>
    <section><title>Section One</title>
        <p>Second paragraph.</p>
        <p>Third paragraph.</p></section>
    <section><title>Section Two</title>
        <p>Fourth paragraph.</p></section>
    <section><title>Section Three</title>
        <p>Fifth <em>paragraph</em>.</p>
        <p>Sixth <em rend='ital'>para</em>.</p>
        <p>Seventh paragraph.</p></section></doc>
```

Example

```
<doc><title>Document Title</title>
    <p>First <em rend='bold'>paragraph</em>.</p>
    <section><title>Section One</title>
        <p>Second paragraph.</p>
        <p>Third paragraph.</p></section>
    <section><title>Section Two</title>
        <p>Fourth paragraph.</p></section>
    <section><title>Section Three</title>
        <p>Fifth <em>paragraph</em>.</p>
        <p>Sixth <em rend='ital'>para</em>.</p>
        <p>Seventh paragraph.</p></section></doc>
```

Example

```
<doc><title>Document Title</title>
    <p>First <em rend='bold'>paragraph</em>.</p>
    <section><title>Section One</title>
        <p>Second paragraph.</p>
        <p>Third paragraph.</p></section>
    <section><title>Section Two</title>
        <p>Fourth paragraph.</p></section>
    <section><title>Section Three</title>
        <p>Fifth <em>paragraph</em>.</p>
        <p>Sixth <em rend='ital'>para</em>.</p>
        <p>Seventh paragraph.</p></section></doc>
```

Example

```
<doc><title>Document Title</title>
    <p>First <em rend='bold'>paragraph</em>.</p>
    <section><title>Section One</title>
        <p>Second paragraph.</p>
        <p>Third paragraph.</p></section>
    <section><title>Section Two</title>
        <p>Fourth paragraph.</p></section>
    <section><title>Section Three</title>
        <p>Fifth <em>paragraph</em>.</p>
        <p>Sixth <em rend='ital'>para</em>.</p>
        <p>Seventh paragraph.</p></section></doc>
```

Example

```
<doc><title>Document Title</title>
    <p>First <em rend='bold'>paragraph</em>.</p>
    <section><title>Section One</title>
        <p>Second paragraph.</p>
        <p>Third paragraph.</p></section>
    <section><title>Section Two</title>
        <p>Fourth paragraph.</p></section>
    <section><title>Section Three</title>
        <p>Fifth <em>paragraph</em>.</p>
        <p>Sixth <em rend='ital'>para</em>.</p>
        <p>Seventh paragraph.</p></section></doc>
```

Example

```
<doc><title>Document Title</title>
    <p>First <em rend='bold'>paragraph</em>.</p>
    <section><title>Section One</title>
        <p>Second paragraph.</p>
        <p>Third paragraph.</p></section>
    <section><title>Section Two</title>
        <p>Fourth paragraph.</p></section>
    <section><title>Section Three</title>
        <p>Fifth <em>paragraph</em>.</p>
        <p>Sixth <em rend='ital'>para</em>.</p>
        <p>Seventh paragraph.</p></section></doc>
```

Example

```
<doc><title>Document Title</title>
    <p>First <em rend='bold'>paragraph</em>.</p>
    <section><title>Section One</title>
        <p>Second paragraph.</p>
        <p>Third paragraph.</p></section>
    <section><title>Section Two</title>
        <p>Fourth paragraph.</p></section>
    <section><title>Section Three</title>
        <p>Fifth <em>paragraph</em>.</p>
        <p>Sixth <em rend='ital'>para</em>.</p>
        <p>Seventh paragraph.</p></section></doc>
```

Example

```
<doc><title>Document Title</title>
    <p>First <em rend='bold'>paragraph</em>.</p>
    <section><title>Section One</title>
        <p>Second paragraph.</p>
        <p>Third paragraph.</p></section>
    <section><title>Section Two</title>
        <p>Fourth paragraph.</p></section>
    <section><title>Section Three</title>
        <p>Fifth <em>paragraph</em>.</p>
        <p>Sixth <em rend='ital'>para</em>.</p>
        <p>Seventh paragraph.</p></section></doc>
```

Example

```
<doc><title>Document Title</title>
    <p>First <em rend='bold'>paragraph</em>.</p>
    <section><title>Section One</title>
        <p>Second paragraph.</p>
        <p>Third paragraph.</p></section>
    <section><title>Section Two</title>
        <p>Fourth paragraph.</p></section>
    <section><title>Section Three</title>
        <p>Fifth <em>paragraph</em>.</p>
        <p>Sixth <em rend='ital'>para</em>.</p>
        <p>Seventh paragraph.</p></section></doc>
```

Example

```
<doc><title>Document Title</title>
    <p>First <em rend='bold'>paragraph</em>.</p>
    <section><title>Section One</title>
        <p>Second paragraph.</p>
        <p>Third paragraph.</p></section>
    <section><title>Section Two</title>
        <p>Fourth paragraph.</p></section>
    <section><title>Section Three</title>
        <p>Fifth <em>paragraph</em>.</p>
        <p>Sixth <em rend='ital'>para</em>.</p>
        <p>Seventh paragraph.</p></section></doc>
```

Example

```
<doc><title>Document Title</title>
    <p>First <em rend='bold'>paragraph</em>.</p>
    <section><title>Section One</title>
        <p>Second paragraph.</p>
        <p>Third paragraph.</p></section>
    <section><title>Section Two</title>
        <p>Fourth paragraph.</p></section>
    <section><title>Section Three</title>
        <p>Fifth <em>paragraph</em>.</p>
        <p>Sixth <em rend='ital'>para</em>.</p>
        <p>Seventh paragraph.</p></section></doc>
```

Example

```
<doc><title>Document Title</title>
    <p>First <em rend='bold'>paragraph</em>.</p>
    <section><title>Section One</title>
        <p>Second paragraph.</p>
        <p>Third paragraph.</p></section>
    <section><title>Section Two</title>
        <p>Fourth paragraph.</p></section>
    <section><title>Section Three</title>
        <p>Fifth <em>paragraph</em>.</p>
        <p>Sixth <em rend='ital'>para</em>.</p>
        <p>Seventh paragraph.</p></section></doc>
```

Example

```
<doc><title>Document Title</title>
    <p>First <em rend='bold'>paragraph</em>.</p>
    <section><title>Section One</title>
        <p>Second paragraph.</p>
        <p>Third paragraph.</p></section>
    <section><title>Section Two</title>
        <p>Fourth paragraph.</p></section>
    <section><title>Section Three</title>
        <p>Fifth <em>paragraph</em>.</p>
        <p>Sixth <em rend='ital'>para</em>.</p>
        <p>Seventh paragraph.</p></section></doc>
```

Example

```
<doc><title>Document Title</title>
    <p>First <em rend='bold'>paragraph</em>.</p>
    <section><title>Section One</title>
        <p>Second paragraph.</p>
        <p>Third paragraph.</p></section>
    <section><title>Section Two</title>
        <p>Fourth paragraph.</p></section>
    <section><title>Section Three</title>
        <p>Fifth <em>paragraph</em>.</p>
        <p>Sixth <em rend='ital'>para</em>.</p>
        <p>Seventh paragraph.</p></section></doc>
```

Example

```
<doc><title>Document Title</title>
    <p>First <em rend='bold'>paragraph</em>.</p>
    <section><title>Section One</title>
        <p>Second paragraph.</p>
        <p>Third paragraph.</p></section>
    <section><title>Section Two</title>
        <p>Fourth paragraph.</p></section>
    <section><title>Section Three</title>
        <p>Fifth <em>paragraph</em>.</p>
        <p>Sixth <em rend='ital'>para</em>.</p>
        <p>Seventh paragraph.</p></section></doc>
```

Example

```
<doc><title>Document Title</title>
    <p>First <em rend='bold'>paragraph</em>.</p>
    <section><title>Section One</title>
        <p>Second paragraph.</p>
        <p>Third paragraph.</p></section>
    <section><title>Section Two</title>
        <p>Fourth paragraph.</p></section>
    <section><title>Section Three</title>
        <p>Fifth <em>paragraph</em>.</p>
        <p>Sixth <em rend='ital'>para</em>.</p>
        <p>Seventh paragraph.</p></section></doc>
```

Example

```
<doc><title>Document Title</title>
    <p>First <em rend='bold'>paragraph</em>.</p>
    <section><title>Section One</title>
        <p>Second paragraph.</p>
        <p>Third paragraph.</p></section>
    <section><title>Section Two</title>
        <p>Fourth paragraph.</p></section>
    <section><title>Section Three</title>
        <p>Fifth <em>paragraph</em>.</p>
        <p>Sixth <em rend='ital'>para</em>.</p>
        <p>Seventh paragraph.</p></section></doc>
```

Example

```
<doc><title>Document Title</title>
    <p>First <em rend='bold'>paragraph</em>.</p>
    <section><title>Section One</title>
        <p>Second paragraph.</p>
        <p>Third paragraph.</p></section>
    <section><title>Section Two</title>
        <p>Fourth paragraph.</p></section>
    <section><title>Section Three</title>
        <p>Fifth <em>paragraph</em>.</p>
        <p>Sixth <em rend='ital'>para</em>.</p>
        <p>Seventh paragraph.</p></section></doc>
```

Example

```
<doc><title>Document Title</title>
    <p>First <em rend='bold'>paragraph</em>.</p>
    <section><title>Section One</title>
        <p>Second paragraph.</p>
        <p>Third paragraph.</p></section>
    <section><title>Section Two</title>
        <p>Fourth paragraph.</p></section>
    <section><title>Section Three</title>
        <p>Fifth <em>paragraph</em>.</p>
        <p>Sixth <em rend='ital'>para</em>.</p>
        <p>Seventh paragraph.</p></section></doc>
```

Example

```
<doc><title>Document Title</title>
    <p>First <em rend='bold'>paragraph</em>.</p>
    <section><title>Section One</title>
        <p>Second paragraph.</p>
        <p>Third paragraph.</p></section>
    <section><title>Section Two</title>
        <p>Fourth paragraph.</p></section>
    <section><title>Section Three</title>
        <p>Fifth <em>paragraph</em>.</p>
        <p>Sixth <em rend='ital'>para</em>.</p>
        <p>Seventh paragraph.</p></section></doc>
```

Example

```
<doc><title>Document Title</title>
    <p>First <em rend='bold'>paragraph</em>.</p>
    <section><title>Section One</title>
        <p>Second paragraph.</p>
        <p>Third paragraph.</p></section>
    <section><title>Section Two</title>
        <p>Fourth paragraph.</p></section>
    <section><title>Section Three</title>
        <p>Fifth <em>paragraph</em>.</p>
        <p>Sixth <em rend='ital'>para</em>.</p>
        <p>Seventh paragraph.</p></section></doc>
```

Example

```
<doc><title>Document Title</title>
    <p>First <em rend='bold'>paragraph</em>.</p>
    <section><title>Section One</title>
        <p>Second paragraph.</p>
        <p>Third paragraph.</p></section>
    <section><title>Section Two</title>
        <p>Fourth paragraph.</p></section>
    <section><title>Section Three</title>
        <p>Fifth <em>paragraph</em>.</p>
        <p>Sixth <em rend='ital'>para</em>.</p>
        <p>Seventh paragraph.</p></section></doc>
```

Example

```
<doc><title>Document Title</title>
    <p>First <em rend='bold'>paragraph</em>.</p>
    <section><title>Section One</title>
        <p>Second paragraph.</p>
        <p>Third paragraph.</p></section>
    <section><title>Section Two</title>
        <p>Fourth paragraph.</p></section>
    <section><title>Section Three</title>
        <p>Fifth <em>paragraph</em>.</p>
        <p>Sixth <em rend='ital'>para</em>.</p>
        <p>Seventh paragraph.</p></section></doc>
```

Example

```
<doc><title>Document Title</title>
    <p>First <em rend='bold'>paragraph</em>.</p>
    <section><title>Section One</title>
        <p>Second paragraph.</p>
        <p>Third paragraph.</p></section>
    <section><title>Section Two</title>
        <p>Fourth paragraph.</p></section>
    <section><title>Section Three</title>
        <p>Fifth <em>paragraph</em>.</p>
        <p>Sixth <em rend='ital'>para</em>.</p>
        <p>Seventh paragraph.</p></section></doc>
```

Example

```
<doc><title>Document Title</title>
    <p>First <em rend='bold'>paragraph</em>.</p>
    <section><title>Section One</title>
        <p>Second paragraph.</p>
        <p>Third paragraph.</p></section>
    <section><title>Section Two</title>
        <p>Fourth paragraph.</p></section>
    <section><title>Section Three</title>
        <p>Fifth <em>paragraph</em>.</p>
        <p>Sixth <em rend='ital'>para</em>.</p>
        <p>Seventh paragraph.</p></section></doc>
```

Example

```
<doc><title>Document Title</title>
    <p>First <em rend='bold'>paragraph</em>.</p>
    <section><title>Section One</title>
        <p>Second paragraph.</p>
        <p>Third paragraph.</p></section>
    <section><title>Section Two</title>
        <p>Fourth paragraph.</p></section>
    <section><title>Section Three</title>
        <p>Fifth <em>paragraph</em>.</p>
        <p>Sixth <em rend='ital'>para</em>.</p>
        <p>Seventh paragraph.</p></section></doc>
```

Example

```
<doc><title>Document Title</title>
    <p>First <em rend='bold'>paragraph</em>.</p>
    <section><title>Section One</title>
        <p>Second paragraph.</p>
        <p>Third paragraph.</p></section>
    <section><title>Section Two</title>
        <p>Fourth paragraph.</p></section>
    <section><title>Section Three</title>
        <p>Fifth <em>paragraph</em>.</p>
        <p>Sixth <em rend='ital'>para</em>.</p>
        <p>Seventh paragraph.</p></section></doc>
```

Example

```
<doc><title>Document Title</title>
    <p>First <em rend='bold'>paragraph</em>.</p>
    <section><title>Section One</title>
        <p>Second paragraph.</p>
        <p>Third paragraph.</p></section>
    <section><title>Section Two</title>
        <p>Fourth paragraph.</p></section>
    <section><title>Section Three</title>
        <p>Fifth <em>paragraph</em>.</p>
        <p>Sixth <em rend='ital'>para</em>.</p>
        <p>Seventh paragraph.</p></section></doc>
```

Example

```
<doc><title>Document Title</title>
    <p>First <em rend='bold'>paragraph</em>.</p>
    <section><title>Section One</title>
        <p>Second paragraph.</p>
        <p>Third paragraph.</p></section>
    <section><title>Section Two</title>
        <p>Fourth paragraph.</p></section>
    <section><title>Section Three</title>
        <p>Fifth <em>paragraph</em>.</p>
        <p>Sixth <em rend='ital'>para</em>.</p>
        <p>Seventh paragraph.</p></section></doc>
```

Example

```
<doc><title>Document Title</title>
    <p>First <em rend='bold'>paragraph</em>.</p>
    <section><title>Section One</title>
        <p>Second paragraph.</p>
        <p>Third paragraph.</p></section>
    <section><title>Section Two</title>
        <p>Fourth paragraph.</p></section>
    <section><title>Section Three</title>
        <p>Fifth <em>paragraph</em>.</p>
        <p>Sixth <em rend='ital'>para</em>.</p>
        <p>Seventh paragraph.</p></section></doc>
```

Example

```
<doc><title>Document Title</title>
    <p>First <em rend='bold'>paragraph</em>.</p>
    <section><title>Section One</title>
        <p>Second paragraph.</p>
        <p>Third paragraph.</p></section>
    <section><title>Section Two</title>
        <p>Fourth paragraph.</p></section>
    <section><title>Section Three</title>
        <p>Fifth <em>paragraph</em>.</p>
        <p>Sixth <em rend='ital'>para</em>.</p>
        <p>Seventh paragraph.</p></section></doc>
```

Example

```
<doc><title>Document Title</title>
    <p>First <em rend='bold'>paragraph</em>.</p>
    <section><title>Section One</title>
        <p>Second paragraph.</p>
                                   </section>


    <section><title>Section Three</title>
        <p>Fifth <em>paragraph</em>.</p>

                                     </section></doc>
```

Example

```
<doc><title>Document Title</title>
    <p>First <em rend='bold'>paragraph</em>.</p>
    <section><title>Section One</title>
        <p>Second paragraph.</p></section>
    <section><title>Section Three</title>
        <p>Fifth <em>paragraph</em>.</p>
                                  </section></doc>
```

Example

```
<doc><title>Document Title</title>
    <p>First <em rend='bold'>paragraph</em>.</p>
    <section><title>Section One</title>
        <p>Second paragraph.</p></section>
    <section><title>Section Three</title>
        <p>Fifth <em>paragraph</em>.</p>
                                </section></doc>
```

Example

```
<doc><title>Document Title</title>
    <p>First <em rend='bold'>paragraph</em>.</p>
    <section><title>Section One</title>
        <p>Second paragraph.</p></section>
    <section><title>Section Three</title>
        <p>Fifth <em>paragraph</em>.</p>
                                    </section></doc>
```

Parameters

- Signature components
  - Ancestor count
  - Preceding-sibling count
  - Self ?
  - Attributes ?
  - Child count
- Repetition count
- Start / middle / end word counts
- Show deletions, signatures, marks

Unimplemented parameters

- Collapse parallel structures

- Single dial
- Target length

- Schema information

- Text abbreviation method

Code

```
<transform
  xmlns='http://www.w3.org/1999/XSL/Transform'
  version='2.0'
  xmlns:tcs='mailto:charlie.hamu@tcs.com'>

<template match='comment() | processing-instruction()'
         mode='#all'>
 <copy />
</template>
```

Passes

```
<template match='/'>
  <variable name='signed'>
    <apply-templates select='node()'   mode='sign'  />
  </variable>
  <variable name='marked'>
    <apply-templates select='$signed'  mode='mark'  />
  </variable>
  <variable name='wrapped'>
    <apply-templates select='$marked'  mode='wrap'  />
  </variable>
  <variable name='needed'>
    <apply-templates select='$wrapped' mode='need'  />
  </variable>
  <variable name='wanted'>
    <apply-templates select='$needed'  mode='want'  />
  </variable>
  <apply-templates   select='$wanted'  mode='prune' />
</template>
```

Sign

```
<template match='*' mode='sign'>
  <copy>
    <attribute name='tcs:signature'
             select='ancestor::*[1]/name(),
                     preceding-sibling::*[0]/name(),
                     self::name(),
                     attribute::*/name(),
                     child::*[0]/name()' />
    <copy-of select='@*' />
    <apply-templates select='node()' mode='sign' />
  </copy>
</template>
```

Mark

```
<template match='*' mode='mark'>
  <copy>
    <if test='not(preceding::*
                      [@tcs:signature
                       = current()/@tcs:signature])'>
      <attribute name='tcs:keep' />
    </if>
    <copy-of select='@*' />
    <apply-templates select='node()' mode='mark' />
  </copy>
</template>
```

Wrap

```
<template match='*' mode='wrap'>
  <copy>
    <if test='not(@tcs:keep)
              and *//@tcs:keep'>
      <attribute name='tcs:keep' />
    </if>
    <copy-of select='@*' />
    <apply-templates select='node()' mode='wrap' />
  </copy>
</template>
```

Need

```
<template match='*' mode='need'>
  <copy>
    <if test='not(@tcs:keep)
              and ../@tcs:keep
              and not(preceding-sibling::*[name()
                                eq current()/name()])
           and not(//*[name() eq current()/../name()
           and not(*[name() eq current()/name()])])'>
      <attribute name='tcs:keep' />
    </if>
    <copy-of select='@*' />
    <apply-templates select='node()' mode='need' />
  </copy>
</template>
```

Want

```
<template match='*' mode='want'>
  <copy>
    <if test='not(@tcs:keep)
              and ../@tcs:keep
              and text()[normalize-space() ne ""]
              and ../text()[normalize-space() ne ""]'>
      <attribute name='tcs:keep' />
    </if>
    <copy-of select='@*' />
    <apply-templates select='node()' mode='want' />
  </copy>
</template>
```

Prune

```
<template match='*' mode='prune'>
  <if test='@tcs:keep'>
    <copy>
      <copy-of select='@* except @tcs:*' />
      <apply-templates select='node()' mode='prune' />
    </copy>
  </if>
</template>

<template match='text()' mode='prune'>
  <value-of select='replace(.,"\.\s.+","...","s")' />
</template>
```

Done

`</transform>`

Discussion

- Other potentially useful parameters?
- Efficient way to mark first signatures?
- Efficient text trimming? Or even bother?
- Efficient Frankenstein? Or even bother?
- Good way to choose the best, not the first?
- Suggested way to read or intuit schema?
- Is this reducible to a single XPath?
- Better done with XQuery?
- How will XSLT 3.0 change this?