An overview of TEI tagging or, Anyone for pizza?

Basic concepts
- The TEI is a modular system, built like a Chicago pizza
- Each module define specific elements and attributes
- Elements are classified semantically and structurally

TEI core modules
- Infrastructure: defines all named element classes and macros
- Common modules for the TEI header and for common elements
- Specialized modules for
  - “book-like” structures of prose, verse, drama
  - speech transcripts
  - dictionaries and terminological lexica

TEI additional modules
- Groups of elements for specialized application areas
- Currently provided:
  - linking and alignment; analysis; feature structures; certainty; physical transcription; textual criticism, names and dates; language corpora; manuscript description....
- Caution! Under Construction!

For example
- TEI Lite (http://www.tei-c.org/Lite/)
  - our guess at what most people want, most of the time realistic for existing texts, and for new document production, e.g. TEI technical documentation
- Modules:
  - core module
  - modules for figures, for linking, for analysis
  - a few omissions

Basic structure(s)
- Every TEI-conformant document comprises a header followed by (at least one) text
- the header contains:
  - mandatory file description
  - optional encoding, profile and revision descriptions
- the header is essential for:
  - bibliographic control and identification
  - resource documentation and processing
Structure of a TEI text

- A text may be unitary or composite
  - a unitary text contains
    - front matter
    - back matter
    - a body
  - in a composite text, the body is a group of texts (or nested groups)

A text usually has divisions

- generic, hierarchic subdivisions
- vanilla or numbered
- **type** attribute
- associated head and trailer elements from the **divtop** class

TEI basic structure

```
<teiHeader>
  <text>
    <!- titlepage, etc here -->
  </text>
</teiHeader>

<teiCorpus>
  <teiHeader>
    <text>
      <!- titlepage, etc here -->
    </text>
  </teiHeader>
  <teiBody>
    <div type='book' n='I' id='JA0100'>
      <head>
        Book I.
      </head>
      <div type='chapter' n='1' id='JA0101'>
        Of writing lives in general,...
        <!- remainder of chapter 1 here -->
      </div>
      <div n='2' id='JA0102'>
        <!- chapter 2 here -->
      </div>
      <!- remainder of book 1 here -->
    </div>
    <div type='book' n='II' id='JA0200'>
      <!- book 2 here -->
    </div>
  </teiBody>
</teiCorpus>
```

A text usually has divisions

- generic, hierarchic subdivisions
- vanilla or numbered
- **type** attribute
- associated head and trailer elements from the **divtop** class

Non-global, but (pervasive) attributes

- **type**
- **target**

TEI global attributes

- Defined in the core module
  - **id** for unique identification
  - **n** for (non-unique) name or number
  - **rend** for rendition (appearance)
  - **lang** for language
- Defined in the linking module
  - **corresp, synch, ana** for specific association types
  - **next, prev** for aggregating fragmented elements
Character Encoding Recommendations

non-normative
extend, using standard entity sets or
transliteration

Document transliteration is not a
System Location

Use Unicode!

Text components (prose base)

- What are divisions composed of?
  - prose is mostly paragraphs (<sp>)
  - verse is mostly lines (<l>), sometimes in hierarchic groups (<lg>)
  - drama is mostly speeches (<sp>) containing <p> or <l> and interspersed with stage directions (<stage>)
- These may be mixed, and may also appear directly within undivided texts.

Verse: an example

```xml
<lg type='haiku'>
<l>Summer grass &mdash</l>
<l>all that's left</l>
<l>of warriors' dreams.</l>
</lg>
```

Drama: an example

```xml
<stage>Enter Barnardo and Francisco, two Sentinels, at several doors</stage>
<sp who='Barnardo'></sp><l>Who's there?</l><sp who='Francisco'></sp><l>Nay, answer me. Stand and unfold yourself.</l><sp who='Barnardo'></sp><l>Long live the king!</l><sp who='Francisco'></sp><l>Barnardo?</l>
```

Texts are not just words...

- … but probably only people know that
  - an encoding may claim to capture
    - just visual salience,
    - just its assumed causes
    - both
  - encoding makes explicit one (or more) sets of interpretations

For example...

```
And this Indenture further witnesseth that the said Walter Shandy, merchant, in consideration of the said intended marriage...
```

```
And this Indenture further witnesseth</hi> that the said <hi rend='italic'>Walter Shandy</hi>, merchant, in consideration of the said intended marriage ...
```
And this Indenture further witnesseth that the said Walter Shandy, merchant, in consideration of the said intended marriage...
My dear Mr. Bennet, said his lady to him one day, have you heard that Netherfield Park is let at last?

Dates, times, numbers

- attributes can be used to quantify <date> and <dateRange> expressions
- similarly, times <time>, <timeRange> and numbers <num>

Correction and Regularization

- <corr> and <sic> for correction (or non-correction)
- <reg> and <orig> for normalization (or the reverse)

Omissions, Deletions, Additions

- <gap> omission by transcriber
- <del> cancellation in source or by editor
- <add> or <supplied> insertion in source or by editor
- <unclear> material uncertain because illegible
- <damage> physical damage to text carrier

The multiple hierarchy problem

- SGML allows only one hierarchy at a time
- Is a document
  - chapter-paragraph-phrase
  - gathering-page-leaf
  - or both?
- discontinuous segments
- links and milestones

Boundary markers

- page, column, and line breaks (<pb>, <cb>, <lb>)
- generic <mileStone>

Names and other referring strings

- The <rs> (referring string) element is used for any kind of name or reference

Omissions, Deletions, Additions

- <gap> omission by transcriber
- <del> cancellation in source or by editor
- <add> or <supplied> insertion in source or by editor
- <unclear> material uncertain because illegible
- <damage> physical damage to text carrier

The multiple hierarchy problem

- SGML allows only one hierarchy at a time
- Is a document
  - chapter-paragraph-phrase
  - gathering-page-leaf
  - or both?
- discontinuous segments
- links and milestones

Boundary markers

- page, column, and line breaks (<pb>, <cb>, <lb>)
- generic <mileStone>

Names and other referring strings

- The <rs> (referring string) element is used for any kind of name or reference

The multiple hierarchy problem

- SGML allows only one hierarchy at a time
- Is a document
  - chapter-paragraph-phrase
  - gathering-page-leaf
  - or both?
- discontinuous segments
- links and milestones

Correction and Regularization

- <corr> and <sic> for correction (or non-correction)
- <reg> and <orig> for normalization (or the reverse)
Some chunks are also phrases

- `<list>` lists of all kinds
- `<note>` notes (authorial or editorial)
- `<figure>` pictures or figures
- `<formula>` formulae
- `<table>` tables
- `<bibl>` bibliographic descriptions

Lists

- use `<list>` for lists of any kind (use `type` attribute to distinguish)
- use `<label>` in two-column lists as alternative to `n` attribute
- may be nested as necessary

Figures and graphics

- The presence of a graphic is indicated by the `<figure>` element
- The title of the graphic is tagged as a `<head>`
- A description of the graphic may be supplied (as a `<figDesc>`) for use by software unable to render the graphic
- The graphic itself is specified by an external link (URL)

for example...

```xml
<list type="xmas">
  <label>For my true love</label>
  <item><list type="bullets">
    <item>three calling birds</item>
    <item>two french hens</item>
    <item>a partridge in a pear tree</item>
  </list></item>
  <label>For Uncle Joe</label>
  <item>socks as usual</item>
</list>
```

for example...

```xml
<figure url="fezz.gif">
  <head>Mr Fezziwig's Ball</head>
  <figdesc>A Cruikshank engraving showing Mr Fezziwig leading a group of revellers.
</figdesc></figure>
```

Tables

- a `<table>` element contains `<row>`s of `<cell>`s
- `spanning` is indicated by `rows` and `cols` attributes
- `role` attribute indicates whether row or column holds data or a label
- embedded tables are permitted
for example...

A three column table

<table>
<thead>
<tr>
<th>Row1</th>
<th>123</th>
<th>4567</th>
</tr>
</thead>
<tbody>
<tr>
<td>Row2</td>
<td>abc</td>
<td>defgh</td>
</tr>
</tbody>
</table>

<table>
    <row cols='3'><cell role='label'>A three column table</cell></row>
    <row><cell role='label'>Row1</cell><cell>123</cell><cell>4567</cell></row>
    <row><cell role='label'>Row2</cell><cell>abc</cell><cell>defgh</cell></row>
</table>

for example...

See for example
<br/>&lt;ref target='REG92'>Regis (1992)&lt;/ref>....

<p>See for example</p>
<br/>&lt;ref target='REG92'>Regis (1992)&lt;/ref>....

Bibliography

- Use simple <bibl> with optional subcomponents:
  - <respStmt> (for any kind of responsibility) or <author>, <editor>, etc.
  - <title> with optional level attribute
  - <imprint> groups publication details
  - <biblScope> adds page references etc.
- Use <listBibl> for list of references

for example...

<ref target='REG92'>Regis (1992)</ref>...

Notes

- Use <note> for notes of any kind (editorial or authorial)
- if in-line, use <place> attribute to specify location
- if out of line, either
  - use <target> attribute to specify attachment point
  - or mark attachment point as a <ref>

The Spoken texts module

- components: <u> <event> <kinesic> <vocal> <pause> <shift>
- contextual information in header <settingDesc> <particDesc>
- facilities for synchronization and timing
**Voice quality and prosody**

- The `<shift>` element is used to mark changes in voice quality
  - `<u who="LB">` <shift feature="loud" new="f"/> Elizabeth</u>
  - `<u who="EB">` Yes</u>
  - `<u who="LB">` <shift/> Come and try this</u>
  - `<u who="LB">` <shift feature="loud" new="f"/> come on</u>

- Other prosodic features may be marked using specific kinds of `<seg>` or entity refs

**Utterances**

- Basic unit of discourse, corresponding to speaker turns
- Optionally grouped into higher-level divisions ( `<div>` ), e.g. to mark discourse function
- Linked by who attribute to `<person>` description in header

**Vocals and events**

- Empty elements are used to mark paralinguistic phenomena

```xml
<u who="Jan">This is just delicious</u>
<event desc='telephone rings'/>
<u who="Kim">I'll get it</u>/
<u who="Tom">I used to <vocal desc='cough'/> smoke a lot</u>/
<u who="Bob"> <vocal desc='sniff'/> he thinks he's tough</u>/
<vocal who="Ann" desc='snorts'/>/
```

**Participant Description**

```xml
<person id="P1" sex="F" age="mid">  
  <birth date='1950-01-12'>
  <date>12 Jan 1950</date>
  <name type="place"> Shropshire, UK </name>
  </birth>
  <firstLang> English </firstLang>
  <langKnown> French </langKnown>
  <residence> Long term resident of Hull </residence>
  <education> University postgraduate </education>
  <occupation> Unknown </occupation>
  <soccstatus source="PEP" code="B2"/>
  </person>
```
Setting Description

Timing
- Pausing
  - use `<pause>` element
- Duration
  - use `dur` attribute
- Overlap
  - use `trans` attribute

Overlap

Have you heard the election results?

<table>
<thead>
<tr>
<th>Have you heard the election results?</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;u id=&quot;A1&quot; who=&quot;A&quot;&gt;Have you heard the&lt;/u&gt;</td>
</tr>
<tr>
<td>&lt;u id=&quot;B1&quot; who=&quot;B&quot; trans=&quot;latching&quot;&gt;the election results? &lt;/u&gt;</td>
</tr>
<tr>
<td>&lt;u id=&quot;A2&quot; who=&quot;A&quot; trans=&quot;pause&quot;&gt;its a disaster&lt;/u&gt;</td>
</tr>
<tr>
<td>&lt;u id=&quot;B2&quot; who=&quot;B&quot; trans=&quot;overlap&quot;&gt;its a miracle&lt;/u&gt;</td>
</tr>
</tbody>
</table>

Not covered here...

- specialised front and back matter
- dictionaries and terminology
- analytic tagging
  - Segmentation
  - interpretations
  - linking
  - the header
  - tags for documentation