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

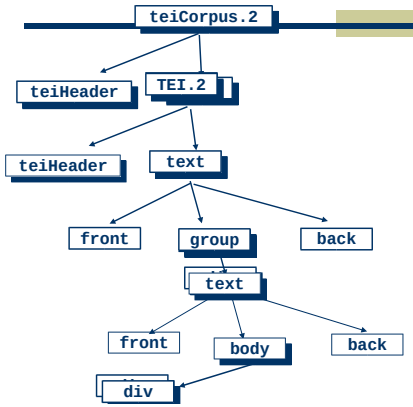
Basic structure(s)

- 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)

TEI basic structures



A text usually has divisions

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

for example...

```
<text>
<front> <!-- titlepage, etc here --> </front>
<body>
  <div type='book' n='I' id='JA0100'>
  <head>Book I.</head>
     <div type='chapter' n='1' id='JA0101'>
       <head>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>
```

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

Non-global, but (pervasive) attributes

- type
- target

Character Encoding Recommendations

- non-normative
- extend, using standard entity sets or transliteration
- document transmitted as Basic Latin with formal *Writing System Declaration*

Use Unicode!

Text components (prose base)

- What are divisions composed of?
  - prose is mostly paragraphs **(<p>)**
  - 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

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

Drama: an example

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

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...

```
<hi rend='gothic'>And this Indenture
further witnesseth</hi> that the said <hi
rend='italic'>Walter Shandy</hi>, merchant,
in consideration of the said intended
marriage ...
```

…or...

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

```
<seg type='formula'>And this Indenture
further witnesseth</seg> that the said
<name rend='italic'>Walter Shandy</name>,
merchant, in consideration of the said
intended marriage ...
```

Who does the work?

- TEI scheme allows for close reading -- and the reverse
- can tag very detailed features of discourse function
- can normalise or simplify (e.g. dates numbers, names)
- … or leave well alone

Core phrase level elements include...

- phrases that are conventionally typographically distinct
- “data-like” (names, numbers, dates, times, addresses)
- editorial intervention (corrections, regularizations, additions, omissions ...)
- cross references and linksb

for example...

```
<head>Of writing lives in general,and
  particularly of <title>Pamela </title>,
  with a word by the bye of <name>Colley
  Cibber</name> and others.</head>
<p>It is a trite but true observation,
  that <q>examples work more forcibly on
  the mind than precepts</q>.…
<p><name>Mr. Joseph Andrews</name>,
  <rs>the hero of our ensuing
  history</rs>, was esteemed to be ...
```

Direct speech

- Use the **who** attribute to show speakers
- Speeches can be nested in other speeches

```
<q who='Wilson'>Spaulding, he came down into
the office just this day eight weeks with
this very paper in his hand, and he
says:&mdash;<q who='Spaulding'>I wish to the
Lord, Mr. Wilson, that I was a red-headed
  man.</q></q>
```

Foreign language phrases

- The **lang** attribute may be attached to any element
- Use <**foreign**> if nothing else is available
- Define each language in <**langUsage**> in header

```
Have you read <title lang='deu'>Die
  Dreigroschenoper </title>?
<mentioned lang='fra'>Savoir-faire</mentioned> is
  French for know-how.
John has real <foreign lang='fra'>savoir-
  faire</foreign>.
```

Names and other referring strings

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

```
<q>My dear <rs type='person'
key='BENM1'>Mr. Bennet</rs>,</q>
said <rs type='person' key='BENM2'>
his lady</rs> to him one day,<q>have
you heard that <rs type='place'
key='NETP1'> Netherfield Park</rs>
is let at last?</q>
```

Dates, times, numbers

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

```
Today is <date>Tuesday 29th</date>.
Today is <date value='1994-11-29'>Tuesday 29th
  </date>.
One afternoon in <date certainty='approx'
  value='1994-11'>late November.</date>.
One afternoon in <dateRange from='1994-11-15'
 to='1994-11-30 exact='to'> late
 November </dateRange>
```

Correction and Regularization

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

```
... for his nose was as sharp as a pen
and  <reg sic="a'">he</reg>
<corr orig='table' ed=Gifford>
babbl'd</corr> of green  <reg
sic='feelds'>fields</reg>
```

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>

```
Diana and <pb ed='ED1' n='475'> Mary
approved the step unreservedly.
Dia<pb ed='ED2' n='483'>na announced
that...
```

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

for example...

```
<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>
```

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...

```
<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		
Row1	123	4567
Row2	abc	defgh

```
<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>
```

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...

```
<p>See for example
<ref target='REG92'>Regis (1992)</ref>....
```

```
<div><head>Bibliography</head>
<listBibl> <bibl id='REG92'>
<author>Ed Regis</author>
<title level=m>Great Mambo Chicken and the Trans-
  Human Experience</title>
<pubPlace>London </pubPlace>
<publisher>Penguin Books</publisher>
<date>1992</date>
<biblscope>pp 144 ff</biblscope></bibl>
  </listBibl></div>
```

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**>

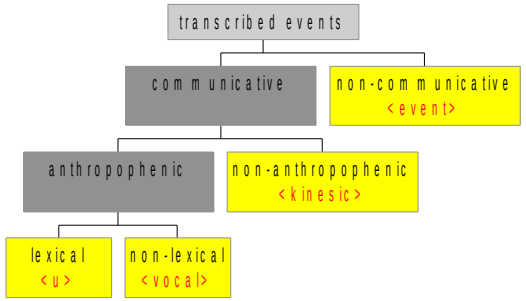
for example...

```
<lg>
<l>The self-same moment I could pray></l>
<l>And from my neck so free</l>
<l>The albatross fell off, and sank</l>
<l id="L213">Like lead into the sea.
   <note type="auth" place="margin">
    The spell begins to break.</note> </l>
</lg>
```

The Spoken texts module

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

Features of speech



Utterances

- Basic unit of discourse, corresponding to speaker turns
- Optionally grouped into higher-level divisions (** **s), 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

```
<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"/>
```

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
<pause/>
<shift feature="loud" new="ff"/>come on</u>
```

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

Another example

```
<u who="MAR">you never <pause/> take this cat for show
and tell <pause dur='5'/> meow meow</u>
<u who="ROS">yeah well I dont want to</u>
<event desc='toy cat has bell in tail which continues
to make a tinkling sound'>
<vocal who="MAR" desc='meows'>
<u who="ROS">because it is so old</u>
<u who="MAR">how <reg orig="bout">about</reg> your cat
<pause/> yours is  new
<kinesic desc='shows Father the cat'></u>
<u who="FAT" trans="pause">that<pause/> darling</u>
<u who="MAR"><s>no mine isnt old</s>
  <s>mine is just um a little dirty</s></u>
```

Participant Description

```
<person id="P1" sex="F" age='mid'>
<birth date='1950-01-12'>
  <date>12 Jan1950</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>
<socecstatus source="PEP" code="B2"/>
</person>
```

Setting Description

```
<settingDesc>
<setting who="P1 P2"><name type="city">Bedford</>
<name type="region">UK: South East</name>
<date value="1989">early spring, 1989</>
<locale>rug of a suburban
home</locale><activity>playing</activity></setting>
<setting who="P3"><name type="city">Bedford</name><name
type="region">UK: South East</name><date value="1989">early spring,
1989</date><locale>at the sink</locale> <activity>washing-
up</activity></setting>
<setting who="P4"><name type="place">London, UK</name>
<time>unknown</time><locale>broadcasting studio</locale>
<activity>radio performance</activity>
</setting></settingDesc>
```

Timing

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

Overlap

Have you heard the
the election results?

```
<u id="A1" who="A">Have you heard the</u>
<u id="B1" who="B" trans="latching">the
election results? </u>
<u id="A2" who="A" trans="pause">its a
disaster</u>
<u id="B2" who="B" trans="overlap">its a
miracle </u>
```

Not covered here...

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