Mr Punch meets the TEI Core

TEI@Oxford

July 2009
The Imaginary Punch Project

• **Punch** is a famous English humorous journal, published regularly between 1841 and 1992: see http://www.punch.co.uk/historyofpunch.html.

• The IPP plans to make available fully marked up texts of the journal, in conjunction with page images...
  - for social historians
  - for librarians
  - for linguists

• How will the TEI help? More specifically, which parts of the TEI will the project use?
Looking at Punch, what do we need to mark up?

- issue information and page number for reference purposes
- "chunks" or divisions of text, which may contain a picture, a poem, some prose, some drama, or a combination
- within the chunks, we can identify formal units such as
  - a picture, a caption
  - stanzas, lines
  - paragraphs
  - speeches and stage-directions
- and more...
TEI tags for the high level structure

We treat each *issue* as a single `<text>` element, and each identifiable chunk within it as a `<div>` element of a particular *type* (e.g. cartoon, verse, prose)

For example, page 1 has two divisions,

```xml
<pb n="1"/>
<div type="cartoon">
  <!-- ... -->
</div>
<div type="poem">
  <head>Progress</head>
  <!-- ... -->
</div>
```

page 2 also has two, of different types:

```xml
<pb n="2"/>
<div type="prose">
  <head>The enchanted castle</head>
  <!-- ... -->
</div>
<div type="snippet">
  <head>Correspondence</head>
  <!-- ... -->
</div>
```
Why divisions rather than pages?

Because a division can start on one page (page 5 for example) and finish on another (page 6)

We use an empty element `<pb>` to mark the boundary between pages, rather than enclosing each page in a `<div type="page">`.

```xml
<pb n="5"/>
<div type="cartoon">
<!-- ... -->
</div>
<div type="review">
  <head>Egypt in Venice</head>
<!-- ... -->
  <pb n="6"/>
<!-- ... -->
</div>
<div type="cartoon">
<!-- ... -->
</div>
<div type="verse">
  <head>Enigma</head>
<!-- ... -->
</div>
<div type="snippets">
<!-- ... -->
</div>
```

The sequence in which divisions appear is rather arbitrary.
Divisions can contain divisions...

<div type="snippets">
  <div type="snippet">
    <!--Curiously....Chancellor-->
  </div>
  <div type="snippet">
    <!--Men for the Antarctic... Canadians-->
  </div>
</div>

- TEI also provides division elements with names that indicate their degree of nesting (<div1>, <div2> etc.) which some people prefer.
- Divisions must always tessellate: once "down" a level, you cannot pop "up" again within the same division.
Global attributes

Some features (potentially) apply to everything:

- identity
- language
- rendition

TEI provides global attributes for these:

- `@xml:id` provides a unique identifier for any element;
- `@n` provides a name or number for any element
- `@xml:lang` specifies the language of any element, using an ISO standard code
- `@rend` and `@rendition` provide ways of specifying the visual appearance (rendition) of any element
More about divisions

Note that in TEI, the heading or headings of a division are part of the division, not separate from it.

<div n="1916-07-02_1147" xml:id="V1470004">
    <head>THE HERE, THERE AND LONDON LETTER.</head>
    <head type="sub">With apologies to <title>The Westminster Gazette</title>.</head>
    <div xml:id="V147000401">
        <head>The Home of the South Saxons.</head>
        <p>Sussex, the county for which Mr. C. B. Fry (who hurt his leg in the
        Lord's centenary match) used to play before he moved to Hampshire, is
        ...
        </p>
    </div>
</div>
Floating text
As mentioned above, `<div>`s must tessellate over the entire text

```
<div1>
  <p> ... </p>
<div2>
  <p> ... </p>
</div2>
<div2>
  <p> ... </p>
</div2>
</div1>
```

is valid BUT

```
<div1>
  <p> ... </p>
<div2>
  <p> ... </p>
</div2>
<div2>
  <p> ... </p>
</div2>
<div1>
  <p> ... </p>
</div1>
```

is not valid.

A special `<floatingText>` element is available for "interruptions"
What are divisions made of?

(apart from other smaller divisions)

- `<head>` (heading)
- `<p>` (paragraph)
- `<sp>` (speech, contains any of the foregoing, also `<stage>` and `<speaker>`)  
- `<list>` (contains `<head>`, `<label>`, `<item>`)  
- `<table>`, (contains `<row>` containing `<cell>`) ...
- `<l>` (verse line) optionally grouped into `<lg>` (line group) stanzas
- `<figure>` (contains `<graphic>`, `<figDesc>`, `<head>`...)
For example....

PUNCH, OR THE LONDON CHARIYARI.—July 1, 1914.

WHEN THE SHIPS COME HOME.

GREECE. "ISN'T IT TIME WE STARTED FIGHTING AGAIN?"
TURKEY. "YES, I DARE SAY. HOW SOON COULD YOU BEGIN?"
GREECE. "OH, IN A FEW WEEKS."
TURKEY. "NO GOOD FOR ME. SHAN'T BE READY TILL THE AUTUMN."
When the ships come home

A man in Turkish dress lounges on a sofa, smoking a cigarette and consulting a book labelled "Naval ledger". Another man, in traditional Greek costume, stands beside him, also reading a notebook.

Greece.

Isn't it time we started fighting again?

Turkey.

Yes, I daresay. How soon could you begin?

Greece.

Oh, in a few weeks.

Turkey.

No good for me. Shan't be ready till the autumn.
For example...

**GREAT ECONOMY EFFECTED BY CO-OPERATION IN ADVERTISEMENT.**

**THE MILITANTS’ TARIFF.**

Mrs. Bangham Sunday, having entered into partnership with the Misses Burnham Blazer, as General Agents of Destruction, begs to inform the public that the firm will be prepared to execute commissions of all kinds, at the shortest notice, on the very moderate terms given below:

<table>
<thead>
<tr>
<th>Service Description</th>
<th>Fee</th>
</tr>
</thead>
<tbody>
<tr>
<td>For breaking windows, per window</td>
<td>0 7 6</td>
</tr>
<tr>
<td>For hurling, kicking, or hitting during service</td>
<td>0 10 6</td>
</tr>
<tr>
<td>For sitting on doorstep of notorious persons, per hour, if fine</td>
<td>0 15 0</td>
</tr>
<tr>
<td>For sitting on doorsteps of notorious persons, per hour, if wet</td>
<td>1 1 0</td>
</tr>
<tr>
<td>For damaging golf greens, per green</td>
<td>2 11 6</td>
</tr>
<tr>
<td>For throwing shoes at magistrates in court, according to size and weight of shoes, per pair</td>
<td>2 2 0</td>
</tr>
<tr>
<td>For beating officials connected with gaols</td>
<td>3 9</td>
</tr>
<tr>
<td>For slashing and hacking valuable pictures, from</td>
<td>7 7 0</td>
</tr>
<tr>
<td>For bombs not intended to explode</td>
<td>8 8 0</td>
</tr>
<tr>
<td>For burning a house, according to value and social position of owner, from</td>
<td>10 0 0</td>
</tr>
<tr>
<td>For insulting exalted Personages, per insult</td>
<td>10 10 0</td>
</tr>
<tr>
<td>For burning down a modern red-brick church</td>
<td>15 15 0</td>
</tr>
<tr>
<td>For burning down a specially valuable and interesting ancient one (eleventh and twelfth centuries extra)</td>
<td>21 0 0</td>
</tr>
<tr>
<td>For bombs warmed to destroy an ordinary church</td>
<td>30 0 0</td>
</tr>
<tr>
<td>For bombs sufficient to reduce really superior buildings, such as Westminster Abbey and St. Paul’s</td>
<td>50 0 0</td>
</tr>
</tbody>
</table>

For disturbing public meetings and the general harassing and annoyance of all peaceable and decent people | No charge.

Bangham Smashy, Burnham Blazer & Co. beg to assure their patrons that all the choppers, hammer, bombs, stones, etc., employed in their business are of the very best quality, and only refined paraffin and wax matches will be used in burnings any building.

Being in a position to offer such exceptional advantages they trust to receive a large measure of support in their elevating and enlightening work.

If none of the above is found suitable to the needs of intending clients, a further list of assorted outrages will be supplied on application.

**LOVE’S LOGIC.**

My happiness is in another’s keeping,
My heart delivered to a maiden’s care,
And she can cast it down or set it leaping
(The latter process is extremely rare)
Ah, would that love indeed had made me blind,
That I might put her image out of mind!

Yet if I looked at her with eyes unsaying
Her voice and laughter would not pass unheard;
I should not be a reasonable being,
I should tremble at her slightest word;
How could I then gain freedom from the spell
Unless I turned completely deaf as well?

So, blind and deaf, I might perhaps recover
A partial peace of mind, but all in vain,
For memories pursue the luckless lover,
And only death can ease him of his pain.

Thrice, having proved that I were better dead,
I think I’ll go and talk to her instead.
<div type="prose">

<head>THE MILITANTS' TARIFF.</head>
<head rend="right">Etna Lodge, W.</head>

Mrs. Bangham Smasher, having entered into partnership with the Misses Burnham Blazer, as General Agents of Destruction, begs to inform the public that the firm will be prepared to execute commissions of all kinds, at the shortest notice, on the very moderate terms given below: --

</p>
<table>
<tr role="label">
  <cell/></tr>
<tr>
  <cell>£</cell>
  <cell>s.</cell>
  <cell>d.</cell>
</tr>
<tr>
  <cell>For breaking windows, per window ...</cell>
  <cell>0</cell>
  <cell>7</cell>
  <cell>6</cell>
</tr>
<tr>
  <cell>For howling, kicking, or biting during service in church, per howl, kick, or bite ...</cell>
  <cell>0</cell>
  <cell>10</cell>
  <cell>6</cell>
</tr>
<tr>
  <cell>For sitting on doorsteps of obnoxious persons, per hour, if fine ...</cell>
  <cell>0</cell>
  <cell>15</cell>
</tr>
</table>
Egypt in Venice (on Page 05) begins with two headings, one in French....

```xml
<div type="prose" xml:lang="en" xml:id="I1914-07-01_05_02">
  <head>Egypt in Venice.</head>
  <head xml:lang="fr" rend="it">"La Légende de Joseph."
  </head>
  <p>Those who know the kind of attractions that the Russian ballet offers in so many of its themes ....</p>
</div>
```
Using the global attributes (2)

Each stanza of the poem on page 10 has a last line which is significantly indented:

```xml
<lg>
  <l>There were eight pretty walkers who went up a hill;</l>
  <l>They were Jessamine, Joseph and Japhet and Jill,</l>
  <l>And Allie and Sally and Tumbledown Bill,</l>
  <l rend="indent">And Farnaby Fullerton Rigby.</l>
</lg>
```
Macrostructure 1

All the issues of *Punch* for one year make up a *volume*. We could regard the volume as a single `<text>`, and each issue as a `<div>` within it. Or we could use the `<group>` element:

```xml
<text xml:id="v147">
  <front>
    <!-- introductory materials for volume 147 here -->
  </front>
  <group>
    <text xml:id="I1914-07-01">
      <body>
        <!-- first issue (1 July) -->
      </body>
    </text>
    <text xml:id="I1914-07-15">
      <body>
        <!-- second issue (15 July) -->
      </body>
    </text>
    <!-- etc... -->
  </group>
  <back>
    <!-- volume index, appendix etc. -->
  </back>
</text>
```
Macrostructure 2

As well as the texts, we have detailed metadata about each volume, and images of its pages. These are the three parts of a canonical TEI document:

```xml
<TEI>
  <teiHeader>
    <!-- required; provides metadata -->
  </teiHeader>
  <facsimile>
    <!-- the text, represented in image form -->
  </facsimile>
  <text>
    <!-- the text, transcribed and marked up -->
  </text>
</TEI>
```
Macrostructure 3

If many such documents are grouped together to form a corpus (rather than a collection), it may be useful to factor out the metadata they have in common:

```xml
<teiCorpus>
  <teiHeader>
    <!-- shared metadata -->
  </teiHeader>
  <TEI>
    <teiHeader>
      <!-- specific metadata -->
    </teiHeader>
    <text>
      <!-- ... -->
    </text>
  </TEI>
  <TEI>
    <teiHeader>
      <!-- specific metadata -->
    </teiHeader>
    <text>
      <!-- ... -->
    </text>
  </TEI>
</teiCorpus>
```
What kinds of metadata?

For *IPP* and for any other comparable project, we will need a place for such information as

- identification of the resource itself ("what is this thing?")
- statements of responsibility ("who did what when?")
- indication of source ("what was this derived from?")
- publication statement ("how is this item distributed and by whom?")
- declaration of encoding practice ("what do the codes we added mean?")

The TEI Header supports all these, and more...
<teiHeader>
 <fileDesc>
  <titleStmt>
   <title>Punch, or the London Charivari, Vol. 147, July 1, 1914</title>
  </titleStmt>
  <publicationStmt>
   <idno type="gutenberg">24357</idno>
   <availability>
    <p>This text is freely available for re-use under US and UK law, consult your local legal restrictions if elsewhere.</p>
   </availability>
  </publicationStmt>
  <sourceDesc>
   <p>This text is a TEI version of a Project Gutenberg text originally located at <ptr target="http://www.gutenberg.org/dirs/2/4/3/5/24357/"/>. As per their license agreement we have removed all references to the PG trademark.</p>
  </sourceDesc>
 </fileDesc>
 <revisionDesc>
  <change when="2008-07-26T23:49:55.968+01:00"/>
 </revisionDesc>
</teiHeader>
Within the elements already introduced, TEI offers plenty of scope for mark-up of smaller components. For example:

- boundaries, such as page, column, or line breaks
- highlighting, emphasis and quotation
- editorial changes such as correction, normalization etc.
- names, numbers, dates, addresses...
- links and cross-references
- notes, annotation, indexing
- graphics
- bibliographic citations
- words and other analyses
By *highlighting* we mean any combination of typographic features (font, size, hue, etc.) which distinguishes the highlighted text from its surroundings. This may be for many reasons...

- to mark foreign, archaic, technical usages
- for emphasis when spoken
- to show something is not part of the text.. (e.g. cross references, titles, headings)
- or is attributed to some other agency inside or outside the text (e.g. direct speech, quotation)

TEI provides both a generic `<hi>` tag and a large number of specific ones...
A few highlighting examples

• <hi> (highlighted: reason unknown or unimportant)

<p>[The rest of this communication is omitted owing to considerations of space.—<hi rend="sc">Ed</hi>.]</p>

• <emph> (emphasized)

<said>'E won't bite yer <emph>if you buy 'im</emph> guv'ner.</said>

• <title> and <foreign>:

<p>À propos</foreign> of Oxford, it is a question whether that extremely amusing book <title>Verdant Green</title> is still much read by freshers.
</p>

• <distinct> (linguistically marked)

But then I remind myself that the Russian ballet is nothing if not <distinct>bizarre</distinct>
Quotation

Quotation marks can similarly be used to set off text for many reasons:

- `<q>` (used if the reason is unknown or unimportant)
- `<said>` (speech or thought)
- `<quote>` (attributed to an external source)
- `<mentioned>` and `<soCalled>` (nuances of narrative status)

```
“I know a lovely tin of potted grouse,” said Celia, and she went off to cut some sandwiches. By twelve
```

```html
<said who="#Celia">I know a lovely tin of potted grouse, </said> said Celia, and she went off to cut some sandwiches.
```
"SoCalled" Quotation

<head>How to utilise the art of <soCalled>suggestion</soCalled></head>
<head>The Doctor, six down at the turn, <soCalled>suggests</soCalled> to his opponent that they are playing croquet, and wins by two and one.</head>
The poet returned to his work. "In tooth and claw," he muttered to himself, "‘In tooth and claw.’"
Editorial intervention

As a simple example, consider:

---

“Excuse me, Sir, but would you like to buy a nice little dawg?”

“No, thanks very much. He looks as though he would bite.”

‘I won’t bite yer if you buy ’im, Guv’ner.’
---

We can:

- use `<orig>` to show that "dawg" is what it says, even though this is a nonstandard spelling
- use `<reg>` to show that "dog" is an editorially-supplied regularisation of what it says
- or provide both within a `<choice>` element to say either is a valid encoding:

...a nice little

```xml
<choice>
  <orig>dawg</orig>
  <reg>dog</reg>
</choice>
```

More (much more) of this kind of thing tomorrow...
Names of persons, places, things...

- `<name>` (a name in the text, contains a proper noun or noun phrase)
- `<rs>` (a general-purpose name or referencing string)
- `<title>` (any form of title)

The `@type` attribute is useful for categorizing these, and they both also have `@key`, `@ref`, and `@nymRef` attributes.
Examples of names

Using *@type* to distinguish personal from geographic names:

```xml
<p>The scene opens at a party given by
<name type="person">Potiphar</name> in
<name type="place">Venice</name>. </p>
```

Using *@key* and *@ref* to de-reference names:

```xml
<p>
  <label>Business done.</label>—The Commons
  still harping on the Budget.
  <name
type="person"
    ref="http://en.wikipedia.org/wiki/Timothy_Michael_Healy">
    Tim Healy</name> enlivened proceedings by vigorous personal attack
  on <q>the most reckless and incapable
  <rs key="LLG">Chancellor of the Exchequer</rs>
  that ever sat on the Treasury Bench.</q>
  <name key="LLG">Lloyd George's</name>
  retort courteous looked forward to with interest.
</p>
```
Dates

• `<date>` contains a date and time in any format
• For processing it is convenient to add a normalized version, using the `@when` attribute
• Uncertain dates and times, and ranges, can be indicated by other attributes: `@notBefore`, `@notAfter`, `@from`, `@to`

*House of Commons, Monday, June 22.*

Great muster of forces on both sides.

House of Commons, `<date when="1914-06-22">Monday, June 22, 1914</date>`.  

`<date notAfter="1914-06-01" notBefore="1914-03-01">Sunday, a month ago,</date>` was hot.
A *cross reference* is a link from one point in a text (the source) to another (the target).

TEI provides generic elements `<ptr>` and `<ref>` for this purpose. If the linking text can be automatically generated use `<ptr>`; otherwise use `<ref>`.

The source is the location of the `<ptr>` or `<ref>`; the target is specified by the `@target` attribute, in the form of a URI reference.

See `<ref target="#Section12">section 12 on page 34</ref>`.

See `<ptr target="#Section12"/>`. 
Bibliographic Citations

TEI provides special elements for bibliographic citations or references:

- `<bibl>` (loosely structured)
- `<biblStruct>` (standard bibliographic structure)
- `<listBibl>` (encloses a bibliography)

These are typically used in preparing bibliographies, or in footnotes. But even in Punch, there are examples.
In Punch, bibliographic citations are usually associated with a quotation from another paper.

```
“\"It was the time when Henry III. was batting with Simon de Montfort and his Barons.\"—Straits Times.
But not at Lord’s, which has only just celebrated its centenary.
```

The `<cit>` element groups the two:

```xml
<cit>
  <quote>It was the time when Henry III. was batting with Simon de Montfort and his Barons.</quote>
  <bibl>
    <title>Straits Times.</title>
  </bibl>
</cit>
```
Embedded notes

Notes, whether appearing in the original source, or added by an editor, can be marked using the `<note>` element. We might use this to add biographical details to the Punch transcriptions:

```
By-the-by, it is denied that Sir Joseph Beecham was in any way responsible for the Government's "Pills for Earthquakes," by which it was hoped to avert the Irish crisis.

---

Sir Joseph Beecham, 1st Baronet
(8 June 1848 - 23 October 1916)...
```

<p>By-the-by, it is denied that Sir <name rend="sc">Joseph Beecham</name> was in any way responsible for the Government's "Pills for Earthquakes," by which it was hoped to avert the Irish crisis.</p>

<note>
Sir Joseph Beecham, 1st Baronet
(8 June 1848 - 23 October 1916)...
</note>

<note> has attributes @place and @resp
Linked notes

Since we have several references to the same person, it might be better to put the notes elsewhere and point to them from the names.

Could also use specialised `<person>` element, in this case. ‘Elsewhere’ can be *anywhere* on the Internet...