Encoding of manuscripts using the TEI

M. J. Driscoll
Arnamagnæan Institute
University of Copenhagen
mjd@hum.ku.dk

TEI Workshop
Azbuky.net
Sofia, Bulgaria
24–26 October 2005
The poem ‘The Old Woman’ first appeared in *Tommy Thumb's Pretty Songbook*, the earliest known book of nursery rhymes, printed ca. 1744. Simple TEI mark-up would look like this:

```xml
<text>
<body>
<div>
<head>The Old Woman.</head>
<lg>
<l>There was an Old Woman,</l>
<l>Liv'd under a Hill,</l>
<l>And if she 'int gone,</l>
<l>She lives there still.</l>
</lg>
</div>
</body>
</text>
```
More information can be added:

```xml
<text>
<body>
<!-- other poems -->
<div type="poem" n="9">
<head>The Old Woman.</head>
<figure>
<figDesc>A rough engraving depicting, in the foreground, an old woman leaning on a fence in front of a small house, and, in the background, a hill, possibly two.</figDesc>
<graphic url="http://www.tomthumb.org/oldwoman.jpg"/>
</figure>
/lg>
<l>There was an <hi rend="capitalize">old woman</hi>,</l>
<l>Liv'd under a <hi rend="capitalize">hill</hi>,</l>
<l>And if &slong;he 'int gone,</l>
<l>She lives there &slong;till.</l>
</lg>
</div>
<!-- other poems -->
</body>
</text>
```
With a simple stylesheet, such as the following:

```css
text {display: block; font: 18pt/20pt "Times New Roman"; padding: 25pt}
head {display: block; font-size: 120%; margin-bottom: 16pt}
l1 {display: block}
*[rend="capitalize"] {text-transform: capitalize}
```

dthis could appear like this:

**The Old Woman.**

There was an Old Woman,
Liv’d under a Hill,
And if she ’int gone,
She lives there still.
What features of the original text might one want to include in a transcription?

- variant letter forms
- orthography
- capitalisation
- word division
- punctuation
- abbreviations
- page lay-out
- additions and deletions
- errors, omissions etc.
Non-standard or ‘exotic’ letter forms can be represented using entities, which may be
given either as numeric entity references in the Universal Character Set developed by
the Unicode Consortium, or using a standardised name which is defined with
reference to the Unicode standard, as in the following example:

```xml
<!ENTITY aelig "&#x00E6;">  
<!ENTITY aeligacute "&#x01FD;">  
<!ENTITY avlig "&#xEF97;">  
<!ENTITY avligacute "&#xEFE7;">  
```

Note that while the first two of these are standard Unicode characters, the latter two
have been defined using the Private Use Area.

With P5 there is available a new mechanism, described in the chapter ‘Representation
of non-standard characters and glyphs’, for dealing with characters which are either
not (yet) available in Unicode, as is often the case when dealing with ancient
languages for which encoding standards do not yet exist, or where one wishes to
distinguish between different allographs of a single character, for example for
purposes of statistical analysis.
By ‘structure’ is meant the division of the work into its constituent parts, by ‘layout’ the arrangement of the text on the page.

The text of the work and the physical object carrying that text have separate structural hierarchies, both of which need ideally to be encoded. For the former the `<div>` element be used for the largest structural divisions in prose texts, with a `type` attribute to specify the nature of the division, ‘chapter’, ‘section’ etc. Paragraphs within these divisions can be tagged using `<p>`. Verse texts can be marked up using the tags `<l>` (for ‘line’) and `<lg>` (for ‘line-group’, i.e. a group of lines functioning as a formal unit), again with a `type` attribute to identify the type of unit, e.g. ‘stanza’, ‘couplet.’ Lines and line-groups can also be numbered and identified using the `n` and `xml:id` attributes. For the structure of the physical document, empty ‘milestone’ elements can be used, `<pb/>`, `<cb/>` and `<lb/>`, for page-, column- and line-boundaries respectively, which can also be numbered and provided with a `xml:id`.
Abbreviations and their expansions can be marked-up using either the `<abbr>` or the `<expan>` element. One may choose either to give the unexpanded abbreviation, tagged `<abbr>`, as in the following example, showing a common Old Icelandic abbreviation for the word hann ('he'), using the `&bar;` (\#x0305;) entity to represent the nasal stroke:

```
hand
```

or use the expanded form, tagged `<expan>`:

```
nann
```

Note that here ‘the abbreviation’ is taken as being the mark or sign used to indicate the suppression of one or more letters, and ‘the expansion’ as being the letters supplied; instead, one may prefer to treat abbreviations and their expansions on a whole-word basis: `<abbr>hand</abbr>` or `<expan>hann</expan>.

One may also provide both the abbreviated and expanded forms, grouped within a `<choice>` element:

```
<h>hand</h><choice>
  <abbr>hand</abbr>
  <expan>hann</expan>
</choice>
```

or:

```
<choice>
  <abbr>h&bar;</abbr>
  <expan>hann</expan>
</choice>
```
Alterations made to the text, whether by the scribe or in some later hand, can be encoded using `<add>` (‘addition’) and `<del>` (‘deletion’); further information may be given as attribute values. The following is an example of a substitution:

Kätlig e&rrot;u þau kuædin<expan>n</expan> <del type="subst" rend="overstrike">&fins;lf</del> <add place="margin" hand="scribe">næ&rrot;ft</add> er koma til iŋgis

Which can be made to display as follows:

Kätlig e₂u þau kuædin ℓeft \næ₂ft/ er koma til iŋgis
The <sic> element can be used to indicate that the reading of the manuscript is erroneous or nonsensical, while <corr> (‘correction’) can be used to provide what in the editor's opinion is the correct reading. These can either be used on their own, i.e. giving only the uncorrected or the corrected reading:

<sic>giorit</sic>
<corr>giorir</corr>

or, as with <abbr> and <expan>, one may choose wrap both inside a <choice> element:

<choice>
  <sic>giorit</sic>
  <corr>giorir</corr>
</choice>
Where a word has been supplied by the editor, the <supplied> tag can be used. It is customary to distinguish between text now illegible or lost through damage but assumed originally to have been in the manuscript (which in some editorial traditions is printed in square brackets), and text assumed to have been inadvertently omitted by the scribe (printed in angle brackets). This distinction is indicated in the mark-up through the use of the reason attribute:

\[lid<\text{supplied } reason=\text{"illegible"}>z</\text{supplied}>\]

\[lid[z]\]

gieck sijdan <\text{supplied } reason=\text{"omitted"}>j burt</\text{supplied}>

gieck sijdan <j burt>

With <supplied> the attribute resp can be used to indicate the person responsible for the conjectural emendation. Where the reading of another witness supports the reconstruction it is also possible to use the source attribute, providing for example the sigil of the other witness:

\[ath \&\text{thorn; eir }<\text{supplied } reason=\text{"omitted" } source=\text{"AM02-152"}>mundu</\text{supplied}> sundr ganga\]

The <supplied> element should only be used when the missing text can be reconstructed with some degree of certainty. When such is not the case <gap> can be used, with both a reason and an extent attribute to indicate the number of characters presumed missing.

<gap reason=\text{"damage" } extent=\text{"7"}/**/
Finally, there is the question of normalisation/regularisation. One can use the `<reg>` element to give regularised forms of variant or archaic spellings, or the `<orig>` element to indicate that a spelling is archaic or non-standard.

In the poem ‘The Old Woman’ there were two non-standard forms, here tagged using the `<orig>` element:

```
There was an Old Woman,
Lived under a hill,
And if she 'int gone,
She lives there still.
```

One could also regularise them and tag them with `<reg>`, or use both elements, grouped within `<choice>` tags:

```
There was an Old Woman,
<choice><orig>Liv'd</orig><reg>Lived</reg></choice>
under a hill,
<choice><orig>'int</orig><reg>isn't</reg></choice>
gone,
<choice><orig>Liv'd</orig><reg>Lived</reg></choice>
She lives there still.
```
Shown here are 8 lines, the beginning of chapter 3, from f. 1v, column b, of the manuscript AM 235 II 4to, an Icelandic vellum of the early 13th century containing a history of the kings of Norway, written probably in the last decade of the 12th century. It contains numerous abbreviations and a number of unusual letter forms. There is also a single error, where the scribe has written the nonsensical ‘er’ (‘is’ or the relative ‘which’) instead of ‘en’ (‘but’).
Jólaaptan Iola aftan

Haraldr haraldr

sat &slong;at

at mat

&thorn;&acute;&thorn;a

kom co<&choice><abbr>&bar;</abbr><expan>m</expan></choice></orig>

Sv&acute;si

&slong;v&acute;&slong;&inodot;

fyrir
<reg>baëdh;</reg>
<orig>baïthon;</orig>
</choice>

<choice>
  <reg>hann</reg>
  <orig>h<choice><abbr>&bar;</abbr><expan>ann</expan></choice></orig>
</choice>
</div>
From a strictly diplomatic text, retaining the line-breaks, variant letter forms, unexpanded abbreviations, and so on of the original:

Iola aftar er haraldr fát at mat. þacō fviði fyr∂ar dvr. oc fendi kí boþ at hí fviði vtt ganga til hif. er kr bræi reθr up∂ þei fendi boþ, oc bar∂farm reθ∂ hif útt er boþ hif haf þí borni inn.
Multiple views to a semi diplomatic text, where the abbreviations have been expanded and the expansions displayed in italics and a obvious error has been corrected:

I ola aftan er haraldr fat
at mat. þacom fíási fyrir
dyr oc fendi konung boþ at hann
févldi útt ganga til hans.
*en konung brazi reiðr up þeim
fendi boþom. oc bar umsami
reið hans útt er boþ hans haf
þi bont inn.
to a semi-normalised text, where abbreviations are expanded silently, the line breaks are not retained, most variant letter forms have been replaced and names have been capitalised:

Iola aftan er Haraldr sat at mat. Þæcom Svíslí svarð dýr oc sendi konungi boþ at hann scvldi útt ganga til hans. *en konungr brazi reiþr úpp þeim sendi boþom. oc bar innsami reiþi hans útt er boþ hans halþi borit inn.
and, finally, to a fully normalised text, where the spelling, punctuation, capitalisation, word division and so on have all been regularised:

**ÓLAAPTAN, er Haraldr sat at mat, þá kom Svási fyrr dyr og sendi konungi bøð at hann skyldi út ganga til hans, *én konungr brásk reiðr við þeim sendiboðum, ok bar inn sami reiði hans út er bøð hans hafði borít inn.**
But this is only the beginning. The TEI also provides mechanisms for associating any kind of semantic or syntactic analysis and interpretation which an encoder might wish to attach to all or part of a text, including familiar linguistic categorisations such as ‘clause’, ‘morpheme’, ‘part-of-speech’ etc. as well as characterisations of narrative structure, such as ‘theme’.