Text Encoding Initiative

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

Text Encoding Initiative

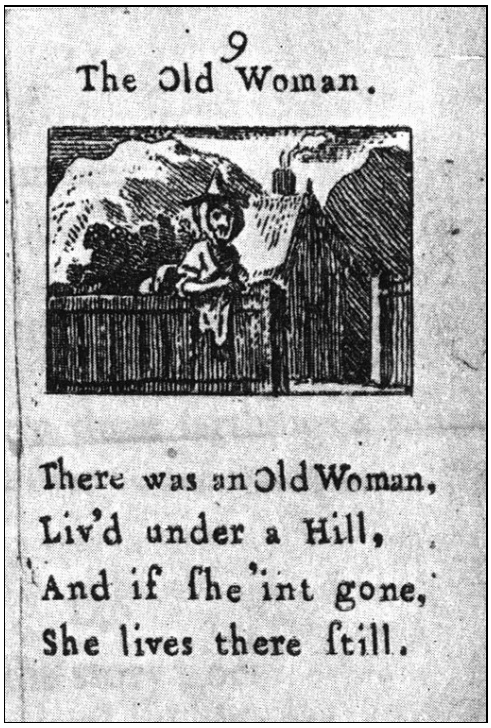
Encoding primary sources

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:

```
<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>
```

Text Encoding Initiative

Encoding primary sources



More information can be added:

```
<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>
```

Text Encoding Initiative

Encoding primary sources

With a simple stylesheet, such as the following:

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

this 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.

Text Encoding Initiative

Levels of transcription

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.

Text Encoding Initiative

Variant letter forms

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:

```
<!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.

Text Encoding Initiative

Structure and layout

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**.

Text Encoding Initiative

Abbreviations

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; (̅) entity to represent the nasal stroke:

```
h<abbr>&bar;</abbr>
```

or use the expanded form, tagged <expan>:

```
h<expan>ann</expan>
```

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 abbrevations and their expansions on a whole-word basis: <abbr>h&bar;</abbr> or <expan>hann</expan>.

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

```
h<choice><abbr>&bar;</abbr><expan>ann</expan></choice>
```

or:

```
<choice><abbr>h&bar;</abbr><expan>hann</expan></choice>
```

Text Encoding Initiative

Additions, deletions and substitutions

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;leſt</del> <add
place="margin" hand="scribe">næ&rrot;ſt</add> er koma til
ijngis
```

Which can be made to display as follows:

Kätlig eʀu þau kuædin*n* ~~ſleſt~~ \næʀſt/ er koma til ijngis

Text Encoding Initiative

Corrections and emendations

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>
```

Text Encoding Initiative

Supplied text

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<supplied reason="illegible">z</supplied>
```

lid[z]

```
gieck sijdan <supplied reason="omitted">j burt</
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 &thorn;eir <supplied reason="omitted" source="AM02-
152">mundu</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="damage" extent="7"/>
```

Text Encoding Initiative

Normalisation

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:

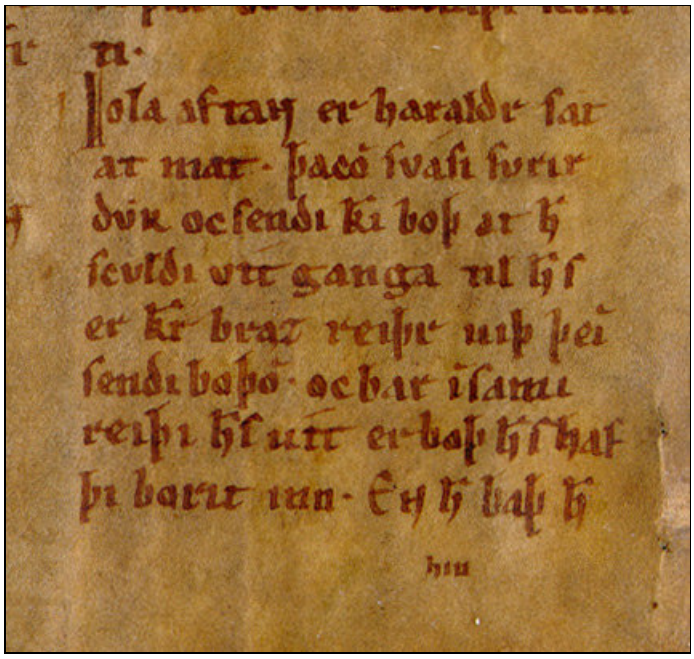
```
<l>There was an Old Woman,</l>
<l><orig>Liv'd</orig> under a hill,</l>
<l>And if she <orig>'int</orig> gone,</l>
<l>She lives there still.</l>
```

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

```
<l>There was an Old Woman,</l>
<l><choice><orig>Liv'd</orig><reg>Lived</reg></choice>
under a hill,</l>
<l>And if she <choice><orig>'int</orig><reg>isn't</reg></
choice> gone,</l>
<l>She lives there still.</l>
```

Text Encoding Initiative

Multi-level mark-up



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’).

Text Encoding Initiative

Multi-level mark-up

```
<div type="chapter" n="3"><p>
<choice>
  <reg><hi rend="3">J</hi>&oacute;laaptan</reg>
  <orig><hi rend="3">I</hi>ola afta&nscap;</orig>
</choice>
<choice>
  <reg>er</reg>
  <orig>er</orig>
</choice>
<name type="person">
<choice>
  <reg>Haraldr</reg>
  <orig>haraldr</orig>
</choice>
</name>
<choice>
  <reg>sat</reg>
  <orig>&slong;at</orig>
</choice>
<lb n="19"/>
<choice>
  <reg>at</reg>
  <orig>at</orig>
</choice>
<choice>
  <reg>mat</reg>
  <orig>mat</orig>
</choice>.
<choice>
  <reg>&thorn;&aacute;</reg>
  <orig>&thorn;a</orig>
</choice>
<choice>
  <reg>kom</reg>
  <orig>co<choice><abbr>&bar;</abbr><expan>m</expan></
choice></orig>
</choice>
<name type="person">
<choice>
  <reg>Sv&aacute;si</reg>
  <orig>&slong;v&aacute;&slong;&inodot;</orig>
</choice>
</name>
<choice>
  <reg>fyrir</reg>
```

```
    <orig>fvr&inodot;r</orig>
  </choice>
  <lb n="20"/>
  <choice>
    <reg>dyrr</reg>
    <orig>d&vdot;&rscap;</orig>
  </choice>
  <choice>
    <reg>ok</reg>
    <orig>oc</orig>
  </choice>
  <choice>
    <reg>sendi</reg>
    <orig>&slong;end&inodot;</orig>
  </choice>
  <choice>
    <reg>konungi</reg>
    <orig>k<choice><abbr>&bar;</abbr><expan>onung</expan></
choice>&inodot;</orig>
  </choice>
  <choice>
    <reg>bo&edh;</reg>
    <orig>bo&thorn;</orig>
  </choice>
  <choice>
    <reg>at</reg>
    <orig>at</orig>
  </choice>
  <choice>
    <reg>hann</reg>
    <orig>h<choice><abbr>&bar;</abbr><expan>ann</expan></
choice></orig>
  </choice>
  <lb n="21"/>
  <choice>
    <reg>skyldi</reg>
    <orig>&slong;cvld&inodot;</orig>
  </choice>
  <choice>
    <reg>&uacute;t</reg>
    <orig>&vacute;tt</orig>
  </choice>
  <choice>
    <reg>ganga</reg>
    <orig>ganga</orig>
  </choice>
  <choice>
    <reg>til</reg>
    <orig>t&inodot;l</orig>
  </choice>
  <choice>
    <reg>hans</reg>
```

```
  <orig>h<choice><abbr>&bar;</abbr><expan>an</expan></
choice>&slong;</orig>
</choice>.
<lb n="22"/>
<choice>
  <corr>
    <choice>
      <reg>en</reg>
      <orig>en</orig>     </choice>
  </corr>   <sic>er</sic>
</choice>
<choice>
  <reg>konungr</reg>
  <orig>k<choice><abbr>&bar;</abbr><expan>onung</expan></
choice>r</orig>
</choice>
<choice>
  <reg>br&aacute;sk</reg>
  <orig>braz</orig>
</choice>
<choice>
  <reg>rei&edh;r</reg>
  <orig>re&inodot;&thorn;r</orig>
</choice>
<choice>
  <reg>vi&edh;</reg>
  <orig>u&inodot;&thorn;</orig>
</choice>
<choice>
  <reg>&thorn;eim</reg>
  <orig>&thorn;e&inodot;<choice><abbr>&bar;</
abbr><expan>m</expan></choice></orig>
</choice>
<lb n="23"/>
<choice>
  <reg>sendibo&edh;um</reg>
  <orig>&slong;end&inodot;
bo&thorn;o<choice><abbr>&bar;</abbr><expan>m</expan></
choice></orig>
</choice>.
<choice>
  <reg>ok</reg>
  <orig>oc</orig>
</choice>
<choice>
  <reg>bar</reg>
  <orig>bar</orig>
</choice>
<choice>
  <reg>inn</reg>
  <orig>&inodot;<choice><abbr>&bar;</abbr><expan>nn</
expan></choice></orig>
</choice>
```

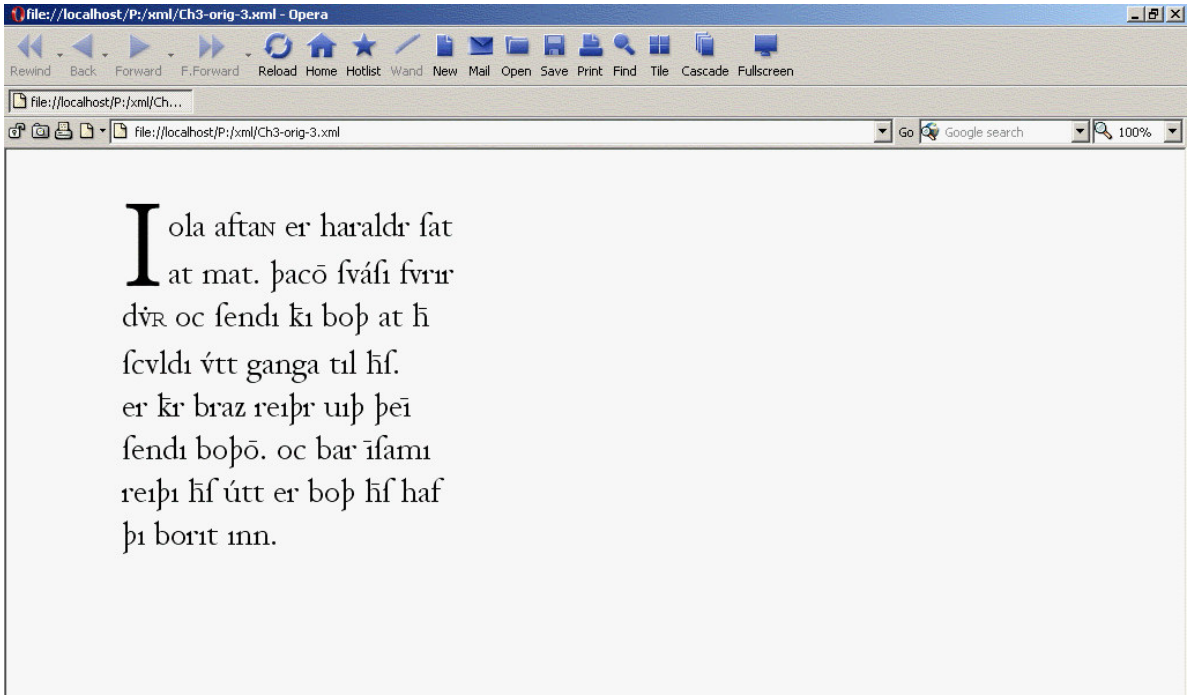
```
<choice>
  <reg>sami</reg>
  <orig>&slong;am&inodot;</orig>
</choice>
<lb n="24"/>
<choice>
  <reg>rei&edh;i</reg>
  <orig>re&inodot;&thorn;&inodot;</orig>
</choice>
<choice>
  <reg>hans</reg>
  <orig>h<choice><abbr>&bar;</abbr><expan>an</expan></
choice>&slong;</orig>
</choice>
<choice>
  <reg>&uacute;t</reg>
  <orig>&uacute;tt</orig>
</choice>
<choice>
  <reg>er</reg>
  <orig>er</orig>
</choice>
<choice>
  <reg>bo&edh;</reg>
  <orig>bo&thorn;</orig>
</choice>
<choice>
  <reg>hans</reg>
  <orig>h<choice><abbr>&bar;</abbr><expan>an</expan></
choice>&slong;</orig>
</choice>
<choice>
  <reg>haf&edh;i</reg>
  <orig>haf<lb n="25"/>&thorn;&inodot;</orig>
</choice>
<choice>
  <reg>borit</reg>
  <orig>bor&inodot;t</orig>
</choice>
<choice>
  <reg>inn</reg>
  <orig>&inodot;nn</orig>
</choice>.
<choice>
  <reg>En</reg>
  <orig>&eunc;&nscap;</orig>
</choice>
<choice>
  <reg>hann</reg>
  <orig>h<choice><abbr>&bar;</abbr><expan>ann</expan></
choice></orig>
</choice>
<choice>
```

```
  <reg>ba&edh;</reg>
  <orig>ba&thorn;</orig>
</choice>
<choice>
  <reg>hann</reg>
  <orig>h<choice><abbr>&bar;</abbr><expan>ann</expan></
choice></orig>
</choice>
<!-- rest of chapter -->
</p>
</div>
```

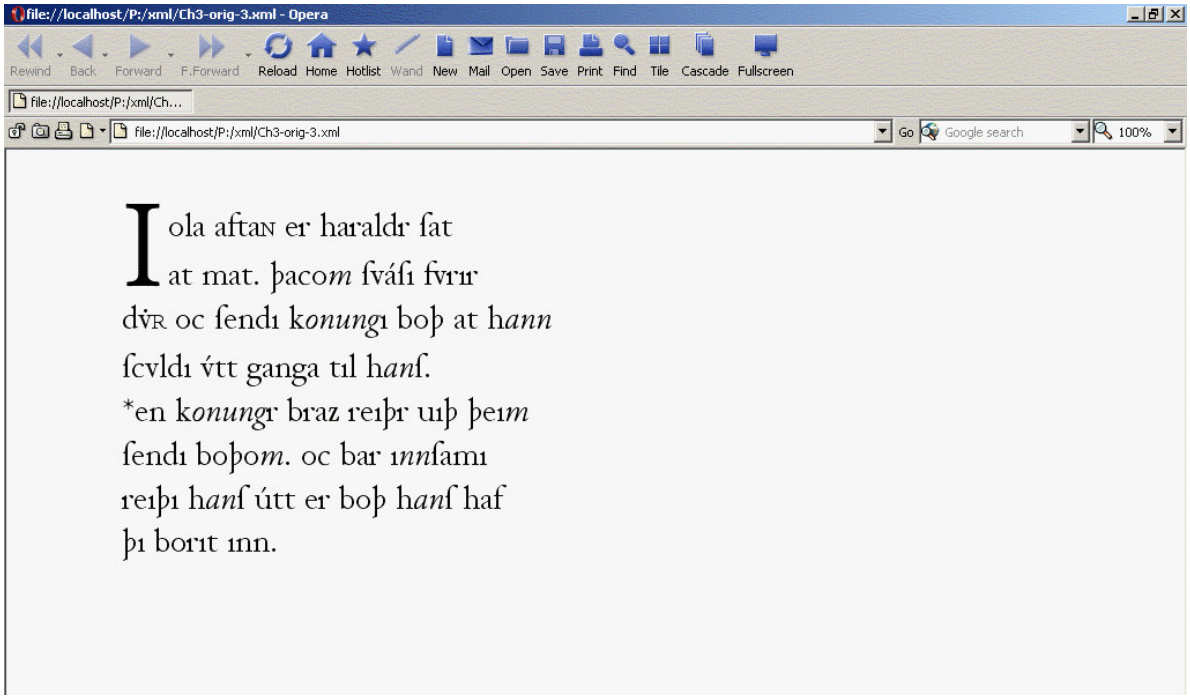
Text Encoding Initiative

Multiple views

From a strictly diplomatic text, retaining the line-breaks, variant letter forms, unexpanded abbreviations, and so on of the original:

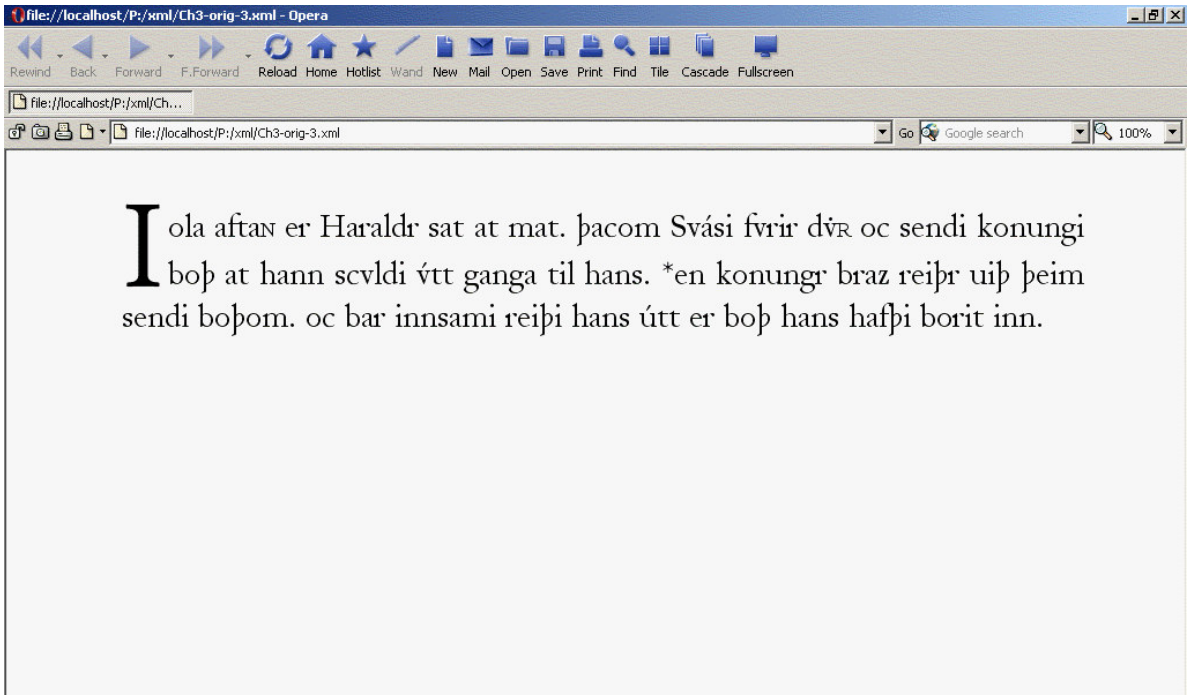


to a semi diplomatic text, where the abbreviations have been expanded and the expansions displayed in italics and a obvious error has been corrected:



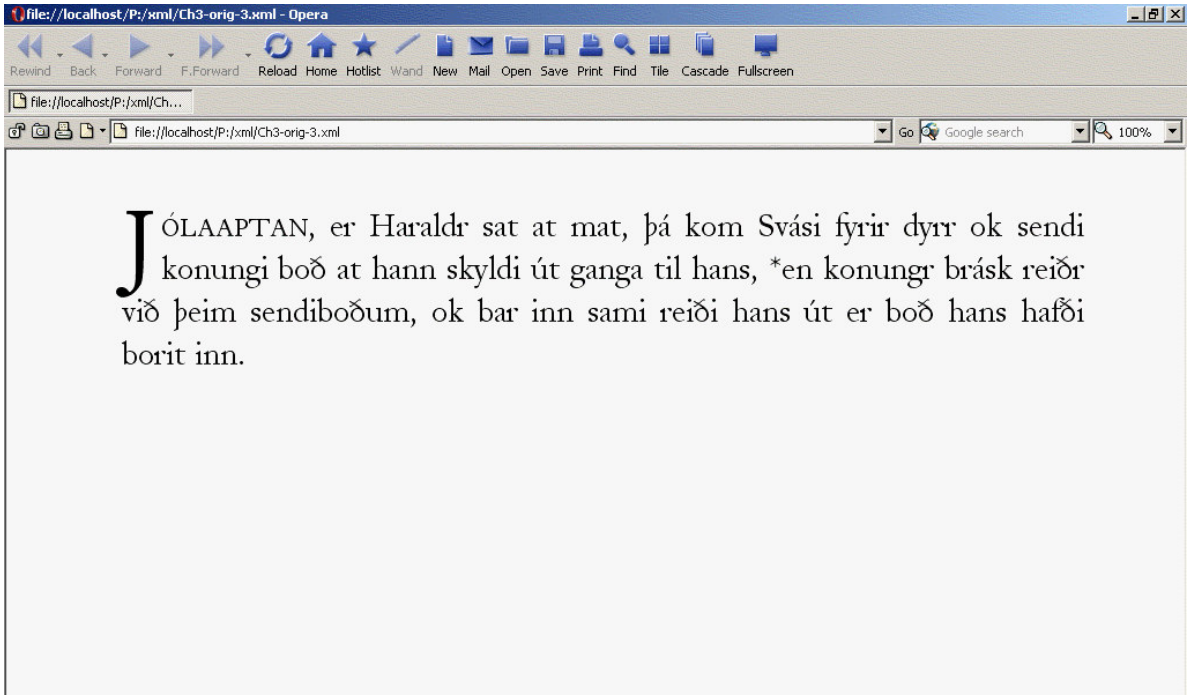
Multiple views

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:



Multiple views

and, finally, to a fully normalised text, where the spelling, punctuation, capitalisation, word division and so on have all been regularised:



Multi-level mark-up

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’.