

## The Text Encoding Initiative



The TEI is an international and interdisciplinary standards project established in 1987 to develop, maintain and promulgate hardware- and software-independent methods for encoding humanities data in electronic form.

The current version of the TEI Guidelines, TEI P5, released in November 2007 (and regularly updated), contains much of potential interest to people working with primary sources such as manuscripts.

It provides mechanisms for encoding in XML:

- A multi-layered transcription of the text
- Information about, and links to, digital facsimiles of the text
- A detailed description of the text-bearing object
- Information about real-world entities associated with the manuscript, i.e. the people, places and organisations named in the text or involved in its production, dissemination and reception

and linking all these together seemlessly



## 1. The transcription of primary sources

The **TEI transcription module** provides mechanisms for encoding not only 'the words on the page', but also such features as:

- Page layout and text structure
- Orthography, capitalisation, word division and punctuation, original as well as normalised Abbreviations and their expansions
- Additions and deletions
- Scribal errors and editorial emendations
- Metrical features
- Variant letter forms



# 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 page.

For the former the <div> element can 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 . Verse texts can be marked up using the tags <|> (for 'line') and <|g> (for 'line-group'), again with a @type attribute to identify the type of unit, e.g. 'stanza', 'couplet'.

For the structure of the physical document, empty 'milestone' elements can be used, <gb/>/s, <pb/>/s, <cb/> and <|b/>lb/>, for gathering-, page-, column- and line-boundaries respectively.



# Abbreviations and their expansion

An abbreviation may be transcribed in two ways:

One may choose to give the unexpanded abbreviation, transcribing it simply as a particular sequence of letters or marks on the page: thus, a 'p with a bar through the descender' or an 'a with a macron'.

One may also interpret or 'expand' the abbreviation, supplying the letter or letters it is seen as standing for.

The TFI proposes two levels of encoding:

- the whole of an abbreviated word and the whole of its expansion can be encoded using <abbr> and <expan>
- the mark or sign used to indicate the suppression of one or more letters, and the letters supplied in the process of expansion can be encoded using <am> and <ex>.

Both levels may be used simultaneously.



# VERSITY OF COPENHAGEN

# A simple example

The Icelandic word 'hann' is frequently written in medieval manuscripts as the letter h with a horizontal stroke or bar (Unicode character 0305).

Depending on editorial policy, this might be represented in any one of the following ways:

<abbr>h&#x305;</abbr>

<expan>hann</expan>

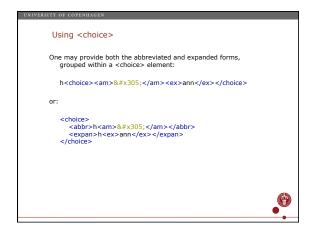
h<am>&#x305;</am>

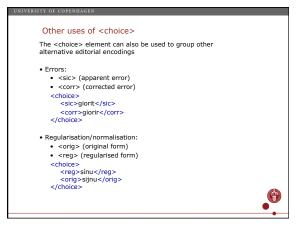
h<ex>ann</ex>

<abbr>h<am>&#x305;</am></abbr>

or <expan>h<ex>ann</ex></expan>



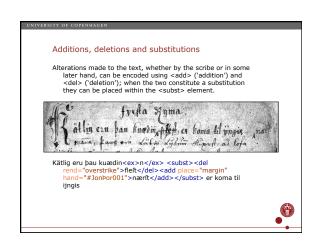


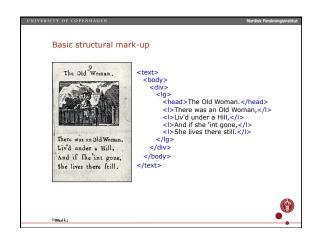


```
Supplied text

Where a word has been supplied by the editor, the <supplied> element can be used. The distinction between text now illegible or lost through damage but assumed originally to have been in the manuscript and text assumed to have been inadvertently omitted by the scribe is indicated through the use of the @reason attribute. Where the reading of another witness supports the reconstruction the @source attribute may be used to identify the other witness.

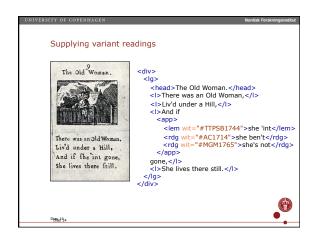
Iid<supplied reason="illegible">z</supplied>
gieck sijdan <supplied reason="omitted">j burt</supplied>
ath þeir <supplied reason="omitted" source="#AM02-152">mundu</supplied> sundr ganga
```

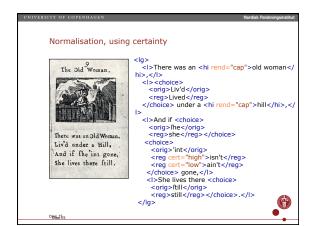


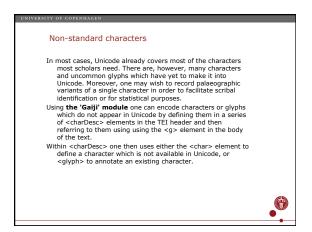










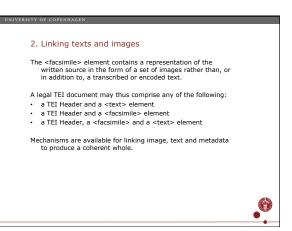


```
Example: Defining a new glyph

A new glyph variant can be defined and assigned to a position in the Unicode Private Use Area (PUA), with a standardised form provided as a fallback:

<chardbecl>
<glyph xmlid="2103">
<glyph xmlid="2103">
<glyph xmlid="2103">
<glyph xmlid="2103">
<glyph xmlid="2103">
<glyph xmlid="2103">
<mapping type="standard">z=/mapping>
<mapping type="standard">z=/mapping>
<flyph>
</chardbecl>

This can now be referenced in the text using the <g> element:
<g ref="#2103">>
It is also possible to provide a default value like this:
<g ref="#2103">z=/g>
```



# The facsimile module The <facsimile> element contains one or more <surface> elements. The <surface> element defines in terms of a rectangular space any written surface. The <zone> element defines a rectangular area within a <surface>. The @facs attribute, available globally, points directly from any element in the text or header to the <surface> or <zone> to which it corresponds. The @start attribute, available on <surface> and <zone>, points to the element containing the transcribed text found within the <surface> or <zone> concerned. In this way one can point from anywhere in a manuscript description or text to the corresponding place on the manuscript page (and vice versa).

```
Aligning text and image

The @facs attributes within the text point to the corresponding <surface> elements, while the @start attributes on <surface> point to the corresponding elements within the text.

<facsimile>
<surface xml:id="image_p1" start="#text_p1">
<sgraphic url="MS101_1r.jpg"/>
</surface>
<surface xml:id="image_p2" start="#text_p2">
<sgraphic url="MS101_1v.jpg"/>
</surface>
<fracsimile>
<text>
<body>
<div>
<pb n="1" xml:id="text_p1" facs="#image_p1"/>
<!-- text of page 1 -->
<pb n="2" xml:id="text_p2" facs="#image_p2"/>
<!-- text of page 2 -->
</div>
</footy>
</footy>
</tobdy>
```

```
Providing coordinates

Coordinates may be provided to locate the surface and any zones defined within it within an abstract space defined by the following attributes:

@ulx gives the x coordinate value for the upper left corner of a rectangular space

@uly gives the y coordinate value for the upper left corner of a rectangular space.

@lrx gives the x coordinate value for the lower right corner of a rectangular space.

@lry gives the y coordinate value for the lower right corner of a rectangular space.
```

```
An example of <facsimile>

<facsimile xml:base="http://www.example.org/MSS/">
<surface xml:id="image_p1" start="#text_p1" ulx="0"
uly="0" |rx="200" |ry="300">
<desc>MS 101, f. 1r</desc>
<graphic mimeType="jpeg" url="MS101-1r.jpg"/>
<zone xml:id="image_p1_det" ulx="20" uly="20" lrx="70"
|ry="70">
<desc>Illuminated initial (detail)</desc>
<graphic mimeType="jpeg" url="MS101-1r-det.jpg"/>
</zone>
</surface>
```

```
3. Information on the text-bearing object

The TEI manuscript description module provides extensive facilities for the description of manuscripts, including:

• The current holding institution

• The intellectual content

• Codicological and palæographical features

• Text layout and illumination/decoration

• Origin and provenance

• Primary and secondary bibliographical references

The TEI <msDesc> element is intended for several different kinds of applications:

• Electronic catalogue record or finding aid

• Metadata component within a digital edition

• Tool for quantitative codicology
```

```
Elements within <msDesc>

<msIdentifier>: groups information uniquely identifying the manuscript, such as holding institution and shelfmark.

<msContents>: provides an itemised list of the intellectual content of the manuscript, with transcriptions of rubrics, incipits, explicits etc., as well as primary bibliographic references.

<pr
```

```
msContents

<msContents

<msCon
```

```
Physical description

The <physDesc> element groups information concerning all physical aspects of the manuscript, its material, size, format, layout, script, decoration, binding, marginalia etc.

Nature of the support

Dimensions of binding, leaves and written area
Foliation, pagination, columnation
Collation (quire structure)
Number of columns, ruled lines, written lines
Presence or absence of catchwords, quire signatures etc.
Text density: lines per page, words/characters per line
Script(s) used; identification of hands
Illumination, decoration, paratextual features
Marginalia
```

```
Another, more data-intensive, example

<supportDesc material="chart">
<support>
<num type="front-flyleaf" value="3"/>
<num type="book-block" value="201"/>
<num type="book-flyleaf" value="3"/>
<dimensions type="leaf">
<dim
```



## 4. Prosopographical and biographical data

TEI P5 provides extensive facilities for the encoding of data pertaining to persons, whether actors in history or those living today.

Such data include:

- · physical characteristics such as sex and eye colour
- cultural characteristics such as socio-economic status and religion
- information on occupation and education, and the events in people's lives such as birth, marriage or appointment to
- the various names by which a person may be known, either in different languages or as they change over time



## The <person> element

There is a specially defined <person> element, within which a large number of sub-elements is available, including:

- <persName>
- <birth> and <death>
- <residence>
- <socecStatus>
- <occupation> and <education>
- <bibl>
- and the generic elements <trait>, <state> and <event>



## Multiple name forms

The <persName> element is repeatable and can, like all TEI elements, take the attribute @xml:lang to indicate the language of the content of the element, thus making it possible to supply name forms in different languages

</persName> 

cypersName xml:lang="da">
<forename sort="2">Arne</forename>
<surname sort="1">Magnussen</surname>

</persName>

# Occupation, education etc.

Information can also be given on occupation, education and socio-economic status. By pre-defining possible values in taxonomies, these can, but need not, appear as empty elements.

<occupation ref="#schol"/>

<education ref="#edu4"/>

<socecStatus scheme="#socecStatus" code="#socec6"/>

The values are defined in the header, e.g.:

<category xml:id="schol"> <desc xml:lang="en">Scholar</desc> <desc xml:lang="is">Fræðimaður</desc>

<desc xml:lang="da">Lærde</desc>

</category>

# 

# A complete <person> element

The <person> element for the scribe Brynjólfur Jónsson á Efstalandi í Öxnadal:

<person sex="1" role="scribe" xml:id="BryJon001">

rison sex-1 Titles Satilex Milliang=18; /erstName writi.ang=18; /forename sort=1">Brynjólfur</forename> /sumame type="patronymic" sort=2">Johnsson</sumame> /persName> /bersName> ore="1600" notAfter="1624">First quarter of the 17th

</person>



